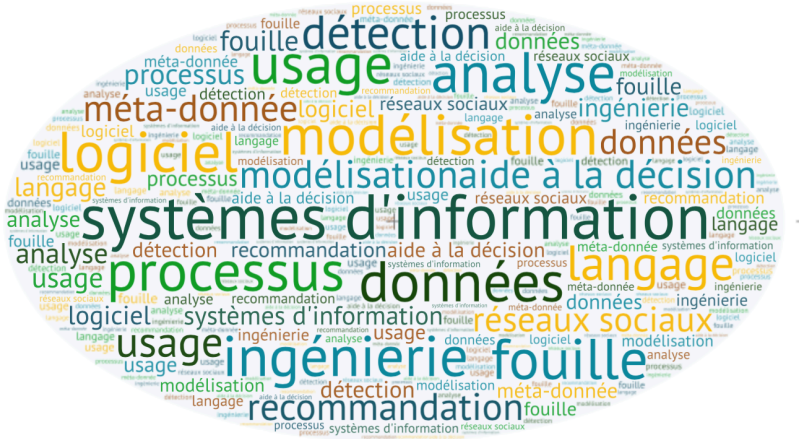
INFORSID 2020

Actes du congrès INFORSID,

38ème édition

prévue à Dijon, 2-4 juin 2020

Editeurs : Sophie Dupuy-Chessa, Thierry Grison



Préface

Depuis plus de 35 ans, la communauté francophone en Systèmes d'Information (SI) organise le congrès INFORSID, lieu privilégié d'échanges sur les avancées de la Recherche, du Développement et de l'Innovation en SI. Après avoir voyagé aux quatre coins du territoire national, INFORSID2020 aurait dû avoir lieu à Dijon du 2 au 4 juin 2020. Ce moment de partage et d'échanges ne pourra pas avoir lieu cette année. Pourtant le contexte actuel nous montre l'importance des SI au quotidien. Il met aussi en exergue le besoin de les faire évoluer pour faire face aux nouvelles pratiques organisationnelles et aux nouveaux défis, comme la sécurité des données, l'exploitation de données massives ou l'aide à la prise de décision. Enfin face au déploiement de solutions technologiques gourmandes en ressources énergétiques, les SI doivent aussi faire face aux défis environnementaux pour proposer des SI éco-responsables.

Dans ce contexte difficile, INFORSID2020 existera néanmoins au travers de la diffusion scientifique et d'échanges à distance. Cette année, le congrès INFORSID a reçu 36 soumissions d'articles dont 6 issues d'articles déjà acceptés dans des conférences ou revues internationales. Les auteurs sont issus de différents pays (France, Tunisie, Belgique, Algérie, Espagne, Suisse, Maroc et Arabie Saoudite, Oman). Dix-neuf articles ont été acceptés : douze dans leur version longue, un dans une version courte et six résumés étendus d'articles déjà publiés dans des conférences ou revues internationales.

Le processus de sélection des articles non publiés dans des conférences internationales s'est déroulé en plusieurs phases. Dans un premier temps, chaque article soumis a été évalué par trois membres du Comité de Programme. Puis, si besoin, des discussions ont eu lieu entre relecteurs pour converger vers un avis par article. Les membres du Conseil du Comité de Programme ont animé ces discussions et rédigé une synthèse des avis. Une réunion à distance du Conseil du Comité de Programme a permis de sélectionner les articles : cinq articles ont été acceptés sous réserves de modifications et un article a été accepté sous condition d'être présenté sous forme courte. Ces articles ont par la suite donné lieu à des relectures et éventuellement un accompagnement pour aboutir à la version présente dans ces actes.

Comme chaque année, les articles couvrent un large panel des problématiques autour des SI: des données aux processus métier, du génie logiciel à l'intelligence artificielle, de l'analyse de solutions existantes à la réalisation de nouveaux systèmes. Ils ont été regroupés thématiquement en six catégories : analyse de l'usage des SI, ingénierie des processus, ingénierie logicielle, aide à la décision et recommandation, analyse de l'information dans les réseaux sociaux et gestion de données complexes.

En complément de la présentation des articles scientifiques, INFORSID2020 aurait dû accueillir un forum Jeunes Chercheuses/Jeunes Chercheurs et trois ateliers (la 4ème édition de l'atelier Systèmes d'Information et de décision, démocratie des organisations et les 2èmes éditions des ateliers sur Evolution des SI- vers des SI pervasifs et sur la Qualité dans l'Internet des objets). Ce n'est bien sûr que partie remise, nous retrouverons ces événements avec plaisir l'année prochaine.

Avant de clore cette préface, je tenais à remercier les différents acteurs de cette édition spéciale d'INFORSID. Sur le plan scientifique, mes remerciements s'adressent à l'ensemble des membres du comité de programme international (France, Luxembourg, Belgique, Algérie, Maroc) et aux membres du conseil du comité. Dans des conditions stressantes, ils ont donné de leur temps et réalisé des retours constructifs dans les meilleurs délais pour que le congrès INFORSID puisse exister cette année. Il ne faut pas oublier les auteurs et les porteurs des ateliers pour leur investissement.

Je tiens également à remercier les membres du bureau de l'association INFORSID, sous la présidence de Franck Ravat, pour m'avoir confié l'organisation scientifique du congrès et pour leur assistance et implication tout au long de cette année.

Enfin je remercie toutes les personnes impliquées dans l'organisation d'INFORSID2020, en particulier Thierry Grison, sans qui le congrès n'aurait pu exister.

Sophie Dupuy-Chessa
Présidente du comité de Programme INFORSID 2020

Le processus de sélection des articles a été géré en utilisant l'outil easychair.

Comités

Le comité de la 38ème édition d'INFORSID est composé par les responsables de l'organisation ainsi que les membres du comité de programme et les membres du conseil du comité de programme.

Comité de programme

Présidente : Sophie Dupuy-Chessa, Univ. Grenoble Alpes, LIG

Membres du Conseil du comité de programme

Cabanac	Guillaume	IRIT	Toulouse
Favre	Cécile	ERIC	Lyon
Marsal de oliveira	Kathia	LAMIH	Valenciennes
Mirallès	André	IRSTEA	Montpellier
Roncancio	Claudia	LIG	Grenoble
Roose	Philippe	LIUPPA	Pau
Souveyet	Carine	CRI	Paris

Membres du Comité de programme

Ahmed-Ouamer	Rachid	Univ. Tizi-Ousou	Algérie
Bastide	Rémi	IRIT	Toulouse
Blay-Fornarino	Mireille	I3S	Nice
Cortès-Cornax	Mario	LIG	Grenoble
Darmont	Jérôme	ERIC	Lyon
Egyed-Zsigmond	Elod	LIRIS	Lyon
Ebersold	Sophie	IRIT	Toulouse
Fredj	Mounia	IAE	Grenoble
Goepp	Virginie	ICube	Strasbourg
Gomez	Paola		
Guedria	Wided	LIST	Luxembourg
Hili	Nicolas	IRT Saint Exupéry	Toulouse
Huchard	Mariane	LIRMM	Montpellier
Idani	Akram	LIG	Grenoble
Kirsch-Pinheiro	Manele	CRI	Paris
Leclercq	Eric	LIB	Dijon
Le Pallec	Xavier	LIFL	Lille
Nègre	Elsa	LAMSADE	Paris
Quinton	Eric	IRSTEA	
Ramadour	Philippe	UPCAM	Aix-Marseille
Savonnet	Marinette	LIB	Dijon
Si-Said Cherifi	Samira	CEDRIC-CNAM	Paris
Sottet	Jean-Sébastien	LIST	Luxembourg
Soulé-Dupuy	Chantal	IRIT	Toulouse

Jean	Stéphane	LIAS	Poitiers
Teste	Olivier	IRIT	Toulouse
Vanderdonckt	Jean	Univ. Catholique Louvain	Belgique
Verjus	Hervé	LISTIC	Annecy

Relecteur externe

Gillet	Annabelle	LIB	Dijon

Comité d'organisation

Président : Thierry Grison, Univ. De Bourgogne, LIB

Membres du comité d'organisation

Journaux	Ludovic	LIB	Dijon
Leclercq	Eric	LIB	Dijon
Savonnet	Marinette	LIB	Dijon

Table des matières

Analyse de l'usage des SI

Ingénierie des processus

Ingénierie logicielle

Aide à la décision et recommandation

Analyse de l'information dans les réseaux sociaux

Gestion de données complexes

Analyse de l'usage des SI

Ce que le numérique fait à l'archéologie et aux archéologues. Un retour d'expériences et un projet de recherche en cours. - *Christophe Tufféry* (article long)

Adoption de l'identifiant chercheur ORCID : le cas des universités toulousaines - *Marie-Dominique Heusse et Guillaume Cabanac* (article long)

Identification de clés pour le succès de projets de gestion informatisée de données environnementales à partir du logiciel Collec-Science - *Eric Quinton, Christine Plumejeaud-Perreau et Sylvie Damy* (article long)

Ce que le numérique fait à l'archéologie et aux archéologues.

Un retour d'expériences et un projet de recherche en cours.

Christophe Tufféry[1]

1. Institut National de Recherches Archéologiques Préventives
121 rue d'Alésia, CS 20007 – F-75685 Paris Cedex 14, France
christophe.tuffery@inrap.fr

RÉSUMÉ. L'article présente le travail en cours d'une thèse d'histoire à CY Cergy Paris Université, en partenariat avec l'Institut National du Patrimoine. Le déploiement en cours et de plus en plus rapide de dispositifs numériques en archéologie demeure faiblement interrogé, en particulier pour les activités de recherche dans cette discipline. Si les nouveaux moyens et les nouvelles pratiques numériques tentent de se rendre accessibles au plus grand nombre, elles laissent aussi de côté une partie des archéologues et révèlent des risques de fracture des collectifs de travail. Depuis une participation observante, l'auteur propose un travail réflexif pour une recherche épistémologique, d'histoire des sciences et techniques et de sociologie des organisations professionnelles de l'archéologie. Les effets étudiés concernent l'archéologie comme discipline et les archéologues comme ensemble de compétences individuelles, de pratiques collectives et d'identités professionnelles. Les matériaux d'étude sont constitués par de nombreuses observations et retours d'expérience depuis plus de dix ans dans le domaine de l'acquisition des données archéologiques de terrain, au sein de l'Institut national de recherches archéologiques préventives (Inrap), et dans plusieurs projets collectifs de recherche pluridisciplinaires.

ABSTRACT. The article presents the ongoing work of a PhD in history at CY Cergy Paris University, in partnership with the Institut National du Patrimoine. The current and increasingly rapid deployment of digital devices in archaeology remains poorly questioned, especially for research activities in this discipline. While the new digital tools and practices attempt to make themselves accessible to the greatest number of people, they also leave out some archaeologists and reveal risks of fracture in work collectives. From an observant participation, the author proposes a reflexive research work in epistemological, history of science and technology and sociology of professional archaeological organisations. The effects studied concern archaeology as a discipline, and archaeologists as a set of individual skills, collective practices and professional identities. The study material is made up of numerous observations and feedback from more than ten years of experience in the field of archaeological field data acquisition at the Institut national de recherches archéologiques préventives (Inrap) and in several collective multidisciplinary research projects.

Mots-clés : numérisation, recherche, archéologie, multidisciplinarité, dispositifs techniques, réflexivité, épistémologie, histoire des sciences et techniques, sociologie des organisations professionnelles, gouvernementalité.

KEYWORDS: digitization, research, archaeology, multidisciplinarity, technical devices, epistemology, reflexivity, history of science and technology, sociology of professional organizations, governmentality

1. Introduction : l'archéologie et le numérique : une histoire déjà ancienne

L'archéologie est la science qui étudie les traces de l'activité des hommes, à travers leurs différentes archives matérielles (structures archéologiques, biens mobiliers et immobiliers, œuvres d'art, archives sonores et visuelles, etc.).

Si l'usage de l'informatique en archéologie a commencé dès les années 1950 avec les travaux de Jean-Claude Gardin (1991), c'est surtout depuis les années 1990 que la production des informations de terrain a commencé à s'appuyer sur des dispositifs numériques. En France, on peut mentionner les systèmes pionniers Syslat (Py, 1991) et Arkeoplan (Buchsenschutz, 1989, Gruel et Buchsenschutz, 1990). Depuis, de nombreuses autres applications ont été développées avec des solutions informatiques très variées : carnets électroniques de terrain des topographes, tableurs, bases de données, SIG, etc. (Desachy, 2008 et 2016). C'est le cas notamment au sein de l'Inrap, le plus grand organisme de recherches archéologiques préventives en France. Depuis 10 ans, nous avons contribué à l'inventaire et à l'étude de plusieurs dizaines des outils numériques utilisés à l'Inrap, en vue de leur harmonisation (Koehler et Tufféry, 2012). Nous avons développé plusieurs applications d'enregistrement de terrain comme l'application EDArc pour la saisie des données de certaines opérations archéologiques de l'Inrap (Tufféry et Augry, 2019) ou encore celles pour les besoins des Projets Collectifs de Recherche sur l'inventaire et l'étude des géoressources siliceuses pour la préhistoire (PCR Réseau de lithothèques) (Fernandes et al., 2013) [1].

Aujourd'hui, l'archéologie fait partie des sciences humaines et sociales qui sont massivement touchées par le déploiement de dispositifs numériques et participe de cet ensemble encore un peu informe que constituent les « humanités numériques » (Mounier, 2012, Le Deuff, 2014).

[1] Les PCR Réseau de lithothèques sont fédérés depuis 2019 dans le cadre d'un Groupement de Recherche (GDR SILEX), cofinancé par le Ministère de la Culture, le CNRS, l'Inrap et Paléotime.

2. Un double point de vue archéologique pour interroger les dispositifs numériques en archéologie

Pour interroger les effets du numérique [2] en archéologie, nous proposons de développer un double point de vue « archéologique ».

D'une part, nous interrogeons le processus de « numérisation du monde » [3] comme une condition de possibilité de la recherche en nous inspirant de la méthode « archéologique » de Michel Foucault. Celle-ci consistait à révéler les soubassements largement impensés des sciences (plus particulièrement les sciences humaines) par l'étude de leur épistémè. Nous proposons d'identifier en particulier comment ce processus de numérisation dans le secteur de la recherche se traduit aujourd'hui pour l'archéologie, en inscrivant cette tentative dans une dimension historiographique de la discipline.

D'autre part, nous proposons à la communauté des archéologues, et à celle des informaticiens qui les accompagnent parfois, des pistes de réflexion en vue d'usages raisonnés des dispositifs numériques qu'ils conçoivent, développent et utilisent au quotidien.

Sans en refaire ici toute l'histoire, que certains font remonter à la machine de Turing (Doueihi, 2013), le numérique s'inscrit d'abord dans l'histoire des sciences et des techniques. A ce titre, divers angles d'étude sont possibles.

Nous délaissons l'approche techniciste du numérique, traditionnellement présentée comme une innovation, pour lui préférer une approche par les usages, centrée sur les objets et ce dont ils témoignent de la réelle adoption des techniques par leurs utilisateurs, à la suite d'Edgerton (2013).

Pour le philosophe Milad Doueihi, le numérique consiste en :

> *une nouvelle manière de fabriquer de la mémoire et de l'interpréter (...) il nous oblige à repenser nos rapports avec ce qui est déjà mémorisé mais également à imaginer de nouvelles façons de préserver et exploiter nos productions purement numériques* (Doueihi, 2013, 54).

Le numérique appliqué aux sciences a d'incontestables avantages, notamment pour le traitement et l'analyse des données à l'aide d'algorithmes qui permettent de répondre à des besoins de représentation ou encore de simulation.

Une approche uniquement techniciste de l'usage du numérique en archéologie ne peut suffire pour en appréhender toutes les manifestations. A la suite de nombreux

[2] Nous utilisons ici le singulier pour désigner un ensemble de techniques qui s'appuient sur le codage d'informations.

[3] L'expression de « numérisation du monde » a été évoquée par le PDG du CNRS Antoine Petit dans une tribune le 26 novembre 2019 dans le journal *Les Echos* (Petit, 2019) à propos du projet de Loi de Programmation Pluriannuelle de la Recherche ». Parmi les objectifs de ce projet de loi et pour en justifier les investissements importants demandés en faveur du numérique, se trouve « *une numérisation du monde au bénéfice du plus grand nombre* ».

auteurs, nous avons choisi de considérer les moyens numériques utilisés non seulement comme des outils techniques (ordinateurs, tablettes, smartphones, écrans, claviers, appareils d'acquisition de données nativement numériques, appareils d'impression sur supports papier ou impression 3D sur divers types de matériaux, casque de réalité immersive, etc.) mais comme un ensemble de dispositifs qui incluent des individus (chercheurs, ingénieurs, techniciens, personnels administratifs, médiateurs, publics divers, etc.) qui les utilisent dans le cadre de leurs pratiques professionnelles, culturelles, touristiques qui elles-mêmes sont encadrées par des codes, des lois, des règlements.

Les dispositifs numériques ne doivent pas être déconnectés de leurs concepteurs, de leurs fabricants, de leurs utilisateurs, ni de leurs usages ni des images des métiers dans lesquels ils sont mis en œuvre. Les dispositifs techniques sont conçus, fabriqués, mis en scène, utilisés, adaptés, remplacés, documentés et institutionnalisés par des organisations sociales, professionnelles ou autres, qui les imposent à leurs membres.

Les dispositifs numériques participent de l'identité des métiers et des individus qui s'en servent. Ils sont embarqués dans les pratiques de leurs utilisateurs. Ils sont des vecteurs en même temps que des marqueurs de la socialisation des individus, où qu'ils se situent dans la chaîne ou plutôt dans leur cycle de vie.

En considérant les outils numériques utilisés en archéologie comme des dispositifs, nous avons souhaité nous interroger sur leurs usages, en particulier dans la recherche effectuée dans cette discipline. C'est l'objet de notre thèse d'histoire et d'épistémologie, engagée à l'automne 2019 à Cergy Paris Université, en partenariat avec l'Institut national de patrimoine.

3. La numérisation comme une condition de possibilité de la recherche archéologique

En prenant l'archéologie comme champ disciplinaire d'étude, nous proposons d'interroger certaines des pratiques de la recherche dans ce domaine, en nous appuyant sur la méthode développée par le philosophe Michel Foucault dans la « période archéologique » de son œuvre au cours des années 1960. Foucault a proposé d'interroger les sciences humaines sur leurs épistémès, c'est-à-dire les conditions de possibilité de leurs savoirs dans une culture et pour une époque donnée, qui se traduisent par des dispositifs (pratiques discursives et non-discursives) qui sont structurés comme des strates, l'analogie entre cette méthode et l'approche stratigraphique des archéologues apparaissant aussitôt.

Pour Foucault, l'épistémè consiste en :

> *l'ensemble des relations pouvant unir, à une époque donnée, les pratiques discursives qui donnent lieu à des figures épistémologiques, à des sciences, éventuellement à des systèmes formalisés ; le mode selon lequel, dans chacune de ces formations discursives, se situent et s'opèrent les passages à l'épistémologisation, à la scientificité, à la formalisation* (Foucault, 1969, 250).

La méthode archéologique foucaldienne vise précisément à identifier les dispositifs en œuvre dans la construction des savoirs par l'analyse de « l'archive », et à en révéler les significations autres que celle que se donnent les savoirs, devenus invisibles en se masquant derrière l'évidence apparente de la raison. Foucault établit ainsi une analogie entre sa méthode et la méthode stratigraphique des archéologues qui consiste à mettre au jour les « archives du sol », à en dégager les relations et à en révéler la véritable signification qui serait masquée par l'évidence de la surface du sol.

Partant de la notion d'épistémè, l'usage du numérique dans la recherche archéologique peut être considéré comme l'une de ses conditions de possibilité.

Pour le chercheur en informatique Gérard Berry, à la suite d'autres auteurs, parfois promus par de grands acteurs économiques du secteur de l'informatique (Hey, T. and *al.*., 2009), le numérique constituerait un changement de paradigme car il s'agirait d' « *une façon de penser et d'agir radicalement différente »* (Berry, 2019, 11).

Avant de pouvoir affirmer que l'usage du numérique constitue un réel changement de paradigme, il conviendrait d'en apporter des preuves multiples et convergentes, en démontrant que le numérique relève bien d'une « révolution scientifique » au sens donnée à cette notion par Thomas Kuhn (2008), à la suite d'Alexandre Koyré. Or, toute « révolution scientifique » intervient parce qu'il y a d'abord une crise dans ce que Kuhn appelle la « science normale », c'est-à-dire les représentations des savants à une époque donnée.

Pour pouvoir affirmer que le numérique constitue une réelle révolution, il manque encore d'une part d'un certain recul qu'impose l'analyse historique qui a besoin d'une certaine profondeur de champ, et d'autre part de cas pratiques analysés, d'expériences étudiées selon des protocoles éprouvés. Sans cela, la « révolution numérique » relève encore d'une sorte de pensée magique, voire d'une idéologie qui ne favorise pas l'esprit critique (Dacheux, 2018). En cela, il est un sujet de prédilection pour les discours mythologiques, largement alimentés par la science-fiction depuis près d'un siècle (Boullier, 2019). A notre sens, la prétendue « révolution numérique » relève d'une série de ce que Gilbert Simondon appelle des « perfectionnements ». Il reste à identifier ce que sont, dans l'histoire du numérique, les « perfectionnements majeurs » et les « perfectionnements mineurs ». Les premiers ressortissent à une véritable rupture dans une lignée technique, par une ou plusieurs mutations orientées par des intentions. Les seconds n'ont que l'apparence des changements et ne présentent aucun bond technique, « *aucune frontière tranchée par rapport à ce faux renouvellement que le commerce exige pour pouvoir présenter un objet récent comme supérieure aux plus anciens* » (Simondon, 1989, 40).

En prenant le domaine de l'archéologie comme objet d'observation, nous pensons utile d'interroger les usages des dispositifs numériques plutôt que de ne présenter que les outils (et de risquer de les survaloriser) numériques (matériels, logiciels, données, etc.), rejoignant en cela les réflexions proposées par l'archéologue britannique Jeremy Huggett depuis une vingtaine d'années (Huggett, 2020).

Comme l'évoque Anne Lehoërff, Vice-Présidente du Conseil National de la Recherche Archéologique (CNRA), le numérique concerne plusieurs des méthodes de l'archéologie depuis le terrain jusqu'aux différentes formes de représentation et de publication des résultats de la recherche archéologique (Lerhoëff, 2019).

Prenons comme premier exemple celui du remplacement progressif des carnets de terrain de l'archéologue par des applications numériques. Traditionnellement, les archéologues utilisent des carnets au format livre de poche avec ou sans élastique, couverture plastifiée ou pas, avec ou sans carreaux, parfois des fiches types pré-imprimées, rangées dans des classeurs d'enregistrement. Dans l'extrême majorité des cas, les carnets de terrain sont propres à l'archéologue. Ils lui appartiennent et lui seul les renseigne. Ses notes, ses commentaires, ses ajouts, ses modifications sur ses notes, ses schémas, ses observations et ses premières interprétations de terrain y sont consignées d'une façon qui lui est particulière et que, parfois, lui seul peut relire et comprendre. L'importance des carnets de terrain pour l'archéologie a été soulignée par Françoise Waquet (2015, 2019). Ces objets matériels sont indétachables de la construction des savoirs. Longtemps, les émotions des chercheurs ne devaient pas être consignées, présentes dans les carnets de terrain, devaient disparaître des publications scientifiques, comme si les chercheurs se devaient de rester totalement extérieurs, neutres, insensibles aux objets de recherche. La dimension sensible et subjective devait être censurée, et n'avait pas le droit d'être assumée ni revendiquée par les chercheurs. Or les carnets de terrain témoignent d'une relation privilégiée entretenue par leurs auteurs avec leurs objets de recherche tout au long du processus d'observation sur le terrain. Ils disent les émotions et la subjectivité des chercheurs. Ces constats ont pu être faits aussi bien pour l'archéologie que pour l'anthropologie, l'ethnologie, l'écologie, la géologie, la géographie et dans toutes les disciplines ou l'observation de terrain est centrale dans la construction des savoirs. Ce n'est que très récemment que la dimension émotionnelle des savoirs a pu commencer à s'exprimer dans les publications scientifiques. Mais la prétention à la neutralité et la mise à distance des émotions par les chercheurs demeure encore largement la norme des publications scientifiques.

Les diverses applications d'enregistrement de terrain sur lesquelles nous avons eu l'occasion de travailler, ont notamment comme principe commun de pouvoir être utilisées par tous les intervenants sur une opération archéologique ou travaillant sur une même thématique de recherche et s'affranchir, en apparence du moins, de toute forme de subjectivité des chercheurs. Avec ces dispositifs, l'archéologue ne peut plus s'exprimer de la même façon ni enregistrer exactement les mêmes types d'informations qu'avec les carnets de terrain traditionnels. Les notes, les commentaires, ne peuvent plus y être écrits ni les schémas y être dessinés de la même façon. Pour autant, l'usage de dispositifs numériques pour l'enregistrement des données d'observation et d'interprétation des archéologues, reste marginal par rapport au volume total des opérations archéologiques réalisée en France chaque année.

La simulation, en plan ou en vue perspective, peut être prise comme un second exemple des domaines de la recherche en archéologie concernés par l'usage de dispositifs numériques, en particulier pour la restitution en 3D de vestiges archéologiques (biens mobiliers), parfois en élévation (biens immobiliers). Dans ce

domaine, le numérique s'est imposé de façon massive depuis une dizaine d'années, soit pour produire des figures statiques, soit pour créer des environnements virtuels interactifs, comme par exemple ceux que proposent les centres d'interprétation des grottes de Lascaux et de Chauvet[4].

Mais, ainsi que l'affirme la chercheuse Sylvie Eusèbe, spécialiste de ce sujet à l'Inrap :

> *il ne faudrait pas céder à la facilité en s'en remettant à la représentation automatique proposée par les machines, se laisser ainsi « déresponsabiliser », et renoncer à l'interprétation de ces images sous prétexte que leur aspect photo-réaliste les rend intelligibles par tous. « La donnée n'est pas la pensée » comme titrait récemment une rencontre interdisciplinaire [5]; la donnée n'est pas le savoir* (Eusèbe, 2019).

Ces divers dispositifs numériques utilisés en archéologie marquent une évolution certaine des outils mais pas des processus intellectuels ni des représentations des archéologues à propos de leur objet d'étude. Pour Huggett, s'il y a un changement de paradigme lié au numérique en archéologie, celui-ci serait à rechercher davantage dans les effets de la production massive de données numériques (*Big Data*) sur les modalités de production et de traitement des savoirs archéologiques, plutôt que dans la définition et l'usage des catégories traditionnelles de ces savoirs (Huggett, 2020).

En nous démarquant d'une opposition dualiste et trop simpliste entre les dispositifs traditionnels et ceux qui relèvent des techniques numériques, nous pensons plus utile de constater que la discipline archéologique conserve une certaine continuité dans ses objectifs et ses grands principes méthodologiques face aux usages grandissants de dispositifs numériques. Ainsi, l'archéologie n'exclut pas aujourd'hui une coexistence entre les modes traditionnels de production et diffusion de la documentation archéologique (carnets de terrain, rapport de fouilles, monographies de sites, synthèses régionales, thématiques, chronologiques, etc.) et leurs équivalents numériques (carnets électroniques de terrain, bases de données ciblées ou généralistes, rapports de fouille en ligne[6], sites cartographiques[7], etc.).

Au-delà des outils numériques eux-mêmes, il nous semble utile d'interroger les effets des dispositifs numériques sur l'archéologie comme discipline et les

[4] Cf. les sites du Ministère de la Culture sur les grottes de Lascaux (https://archeologie.culture.fr/lascaux/fr) et de Chauvet (https://archeologie.culture.fr/chauvet/fr), complété tout récemment par le projet « *Chauvet, à l'aube de l'art* » produit par le Syndicat mixte de la Grotte Chauvet 2 et l'Institut Art & Culture de Google, associés à AURA-Cinéma (https://artsandculture.google.com/project/chauvet-cave?hl=fr)

[5] http://obvil.paris-sorbonne.fr/actualite/la-donnee-nest-pas-la-pensee/jeu-14122017-0000

[6] Plate-forme Dolia de l'Inrap (http://dolia.inrap.fr/) et site des Documents d'Archéologie Préventive de l'Inrap (https://www.inrap.fr/dap/accueil)

[7] Ex. Atlas de l'Äge du Fer de l'UMR AOROC (CNRS-Ecole Normale Supérieure (https://www.chronocarto.eu/gcserver/patlas)

archéologues dans leurs pratiques de quotidiennes et dans le fonctionnement de leurs collectifs professionnels de travail. En quoi l'usage de ces dispositifs ne concerne pas seulement les habitudes de travail mais aboutit à une refonte des principes mêmes de la discipline ? En quoi l'évolution de ces dispositifs modifie-t-il le comportement, les pratiques, les discours et jusqu'aux institutions mêmes de l'archéologie ? La réponse à ces questions est peut-être, là aussi, à rechercher du côté de l'œuvre de Michel Foucault.

4. Le numérique comme entreprise de gouvernementalité technique

Comme tout dispositif technique, ceux du domaine du numérique relèvent d'une entreprise de gouvernementalité, dans le sens que Foucault donna à cette notion qui désigne d'une part « *les institutions, les procédures, analyses et réflexions, les calculs et les tactiques qui permettent d'exercer cette forme bien spécifique, quoique très complexe de pouvoir qui a pour cible principale la population, pour forme majeure de savoir l'économie politique, pour instrument essentiel les dispositifs de sécurité »* et d'autre part *« la tendance, la ligne de force qui, dans tout l'Occident, n'a pas cessé de conduire, et depuis fort longtemps, vers la prééminence de ce type de "gouvernement" sur tous les autres : souveraineté, discipline, et qui a amené, d'une part, le développement de toute une série d'appareils spécifiques de gouvernement, et, d'autre part, le développement de toute une série de savoirs* (Foucault, 2004, 111-112).

Dans *L'art de ne pas être trop gouverné* (2019), le philosophe Jean-Claude Monod évoque ce qu'il appelle le « panoptique numérique » en référence au panoptique de Michel Foucault, un dispositif de surveillance permanente des individus qui eux, ne peuvent savoir quand ni par qui ils sont surveillés.

De leur côté, Antoinette Rouvroy et Thomas Berns proposent une extension de la notion de gouvernementalité, telle que définie par Michel Foucault, avec celle de « gouvernementalité algorithmique » (Rouvroy et Berns, 2009). Parmi les formes que prend cette dernière, les auteurs soulignent l'usage des normes.

En archéologie, les pratiques des chercheurs se voient imposer de plus en plus l'usage de normes diverses, certaines formelles, d'autres plus informelles :

- normes des formats numériques à utiliser et qui doivent assurer l'interopérabilité technique et sémantique des données, par l'usage de formats ouverts,
- normes de « bonnes pratiques » qui visent à distinguer celles définies comme « bonnes » et donc « normales » par opposition à toutes les autres qui se retrouvent de fait qualifiées de pratiques « anormales »,
- normes du développement d'applications et de programmes informatiques qui proviennent non pas du domaine de l'archéologie mais de celui de l'informatique et des sciences de l'information (modélisation conceptuelle, génie logiciel, documentation du code, documentation développeur, documentation utilisateur, remontées et corrections de bugs, versionnement des applications, etc.).

Parmi les nouvelles normes qui vont progressivement s'imposer à la recherche, en archéologie comme dans tout autre discipline, se trouvent les plans de gestion des données ou PGD[8]. Ces plans imposent que les données de recherche produites sur financements publics, du moins certains d'entre eux, s'inscrivent dans un cadre descriptif prédéfini, permettant la réutilisation de ces données selon les principes FAIR [9]. Sans vouloir contester fondamentalement leur intérêt pour les divers acteurs concernés de la recherche, de telles pratiques normalisées peuvent être considérées comme les moyens d'une gouvernementalité technique et pas seulement comme ceux d'une exigence accrue de scientificité.

Toutes ces normes, très codifiées, « disciplinarisent » les pratiques de recherche auxquelles elles s'imposent. Elles le font notamment par leurs effets sur les environnements dans lesquels les chercheurs exercent leur activité. Ainsi, ces normes imposent aux pratiques de recherche et aux chercheurs, de nouvelles conditions de possibilité et de nouveaux horizons affirmés comme incontournables.

Cette généralisation de la normalisation, toujours présentée au service du déploiement de la modernité et du mythe du progrès, a été dénoncée par sociologue français Jacques Ellul dans l'un de ses ouvrages majeurs :

> *Il faut créer pour tout des normes, car la normalisation des données constitutives de la société, de l'être humain, permet seule l'application intégrale des techniques et en même temps permet seule l'universalisation* (Ellul, 2012, 408).

Pour C. Dubar, les changements liés aux normes introduisent des perturbations dans les identités sociales et professionnelles :

Le changement de normes, de modèles, de terminologie provoque une déstabilisation des repères, des appellations, des systèmes symboliques antérieurs. Cette dimension, même si elle est complexe et cachée, touche une question cruciale : celle de subjectivité, du fonctionnement psychique et des formes d'individualité ainsi mises en question (Dubar, 2000).

Parmi les autres enjeux du numérique, se trouve l'ambition d'une accélération de la temporalité dans laquelle s'exerce le travail des chercheurs. Le scénario idéal veut que les chercheurs profitent des outils numériques pour travailler plus vite, pour dégager du temps pour multiplier leurs activités, leur « productivité », devenue une notions envahissant le quotidien des chercheurs et l'aune de laquelle ils sont « évalués » (Dejours, 2003). Cette accélération des pratiques de recherche s'appuie notamment sur l'accroissement des capacités de calcul et de fréquences des

[8] Pour l'archéologie, il existe le PGD élaboré dans le cadre du groupe de travail du consortium Mémoires des archéologues et des sites archéologiques (MASA) : https://masa.hypotheses.org/category/plan-de-gestion-de-donnees

[9] L'acronyme FAIR pour *Findable, Accessible, Interoperable, Reusable*, traduit en français par Facile à trouver, Accessible, Interopérable et Réutilisable, correspond aux quatre principes que doivent respecter toutes les données ouvertes, notamment celles produites dans le cadre de la science ouverte : https://fr.wikipedia.org/wiki/Fair_data

processeurs, mais aussi sur l'interopérabilité entre applications numériques ou encore entre outils ou objets dits « connectés ». Les communications entre individus sont plus rapides, les fichiers sont échangés plus vite, les données sont relues directement, etc. bref, tous ces principes sont au service du processus d'accélération dont le philosophe et sociologue allemand Hartmut Rosa a démontré les effets délétères :

> *Les forces d'accélération de la société contemporaine, dans le passage du XXe au XXIe siècle, engendrent une redéfinition du rapport à soi-même, sur le plan individuel et collectif, c'est-à-dire des formes dominantes d'identité, de même que des formes de l'activité ou de l'organisation politique* (Rosa, 2010, 39).

H. Rosa a récemment proposé la thèse que la numérisation participe du projet de « rendre le monde disponible » à tous, humains et machines, en tout lieu et à tout moment. Or, pour cet auteur, il y a urgence à ne pas prolonger cette entreprise mais à l'inverser pour permettre à l'humanité de trouver une autre relation avec la technique et pouvoir de nouveau entrer en résonnance avec elle (Rosa, 2020).

Tenter de révéler ainsi ce que cache le projet de numérisation de la recherche archéologique, c'est emprunter à la méthode archéologique foucaldienne son principal objectif qui est de faire ressortir l'épistémè des savoirs archéologiques pour l'époque actuelle. Cette épistémè est soutendue par une nouvelle conception du monde qui s'appuie sur des critères renouvelés de véridiction des pratiques actuelles de la recherche, qui témoignent de l'établissement du régime de vérité de ces pratiques. Le processus de numérisation du monde et de celui de la recherche en particulier, entraîne une reconfiguration des métiers, des pratiques et des identités professionnelles des chercheurs. Or, si comme l'affirme le PDG du CNRS, le processus en cours de numérisation de la recherche a l'ambition d'être au bénéfice du plus grand nombre, cela signifie, en creux, qu'il ne cherche pas à l'être au bénéfice de tous. Pourquoi cette ambition limitée ? Cette entreprise considère-t-elle possible de laisser certains chercheurs de côté ?

5. Le numérique au bénéfice du plus grand nombre mais pas de tous

Certes, pour une partie des archéologues, comme pour de nombreux autres chercheurs, l'usage de dispositifs numériques ouvre à des possibilités nouvelles pour faire évoluer leurs pratiques de recherche. Mais pour d'autres, la mise en œuvre de ces dispositifs les conduit à des évolutions contraintes et à dans certains cas, à des adaptations forcées à ces dispositifs (Tufféry, 2019).

Ces changements de pratiques imposent aux archéologues de se former à divers dispositifs numériques (tablettes, appareils de topographie et de photographie numérique, outils d'étude non invasive des sols et des matériaux, etc.), mais aussi parfois à des langages et des méthodes propres au domaine de l'informatique et, prochainement, à de nouveaux modes de production et de diffusion de leurs connaissances comme les ontologies et leurs langages de description (RDF, XML, SKOS, etc.).

Plus encore, ce sont les pratiques des archéologues en matière de de stockage, d'archivage et de conservation des données archéologiques qui sont affectées par le déploiement d'outils numériques. Dans ces domaines, les archéologues sont invités à faire évoluer leurs pratiques et à faire preuve d'une vigilance accrue par rapport aux données qu'ils produisent, à multiplier les procédures de sauvegarde et donc les copies de leurs données, quitte à utiliser parfois des moyens personnels (stockage sur des disques durs externes, des clés USB, dans le « nuage » sur leur espace personnel sur des plateformes commerciales, etc.). En poussant à ce type de pratiques, l'usage des dispositifs numériques se traduit par des copies multiples des données sur des supports variés, personnels et professionnels, mais sans le souci d'en garantir pour autant l'intégrité, l'accessibilité, la sécurité ni la réutilisabilité.

Le sociologue Antonio Casilli (2010) a montré que l'ordinateur est l'un des rares dispositifs techniques à modifier les trois espaces de la sphère domestique à la fois : espace physique (répartition, aménagement et usages des pièces, agencement des objets), espace technologique (outils, instruments utilisés pour répondre à des besoins), espace social (liens de sociabilité entre les individus par lesquels ils se construisent, s'individualisent, s'identifient). Il en est de même pour les dispositifs numériques qui modifient les trois espaces de la sphère professionnelle : l'espace physique (bureaux, laboratoire, terrain, lieux de rencontre entre individus, etc.), l'espace technologique (ordinateurs, tablettes, smartphones, serveurs, réseaux, etc.), espace social (liens de sociabilité par lesquels les individus s'affirment et construisent leur identité professionnelle).

Les moyens numériques s'insinuent partout dans les lieux de la recherche, ils créent une nouvelle territorialisation, ils imposent de nouveaux lieux, de nouvelles locaux (ex. des salles dédiée exclusivement aux serveurs informatiques), ils impliquent de mettre en place de nouvelles infrastructures (ex. réseaux informatiques spécifiques à la recherche comme le réseau RENATER), d'installer de nouveaux équipements (matériels informatiques en tous genres, armoires de brassage des réseaux, etc.). Cette reterritorialisation de la recherche participe d'une nouvelle configuration des « lieux de savoirs » pour reprendre le titre de l'ouvrage dirigé par Christian Jacob (2007, 2011). En même temps qu'elle se déploie, la numérisation bénéficie de la miniaturisation des moyens techniques sur lesquels elle s'appuie : miniaturisation des circuit, remplacement des ordinateurs par des tablettes et des smartphones plus petits et plus légers. De façon paradoxale, la diminution de l'encombrement des moyens techniques de la numérisation pour un même niveau de capacité de calcul et de stockage, profite non pas à une diminution de l'occupation des lieux des dispositifs numériques mais au contraire à son entreprise d'extension dans l'espace. Cette conquête des espaces de travail des chercheurs par le numérique montre bien que la « dématérialisation » des savoirs modernes passe par une extension des espaces occupés par les moyens matériels nécessaires à cette dématérialisation.

Le numérique se traduit par des changements multiples et plus ou moins profonds sur ce que le sociologue Claude Dubar appelle les identités professionnelles, qui relèvent de deux catégories d'identification, « externes » (*pour autrui*) et « internes » (*pour soi*) (Dubar, 1991). Ces modifications concernent aussi bien les façons d'apprendre, que les compétences déjà acquises et d'autres à

acquérir, ou encore les savoir-faire, les manières de faire, qui dépassent les simples compétences. Les savoirs et les savoir-faire voient leur valeur évoluer dans les dimensions scientifique, économique, symbolique, etc.

Le numérique intervient dans la « mise en scène » de la vie quotidienne pour reprendre la notion du sociologue Erving Goffman (1973). Pour Goffman, les individus masquent leurs comportements par des éléments scéniques et matériels de leur environnement de travail qui constituent le « décor » (mobilier, objets, accessoires, etc.) dont le sociologue français Pierre Bourdieu a montré la fonction sociale de représentation et de distinction (Bourdieu, 1979). Les éléments matériels constituent la perspective « technique » de la mise en scène quotidienne de soi au travail. A côté de cette dimension matérielle, il en existe une autre, que Goffman appelle la « façade personnelle » et qui regroupe les attributs confondus avec la personne (ses façons de faire, de se mouvoir, de se saisir d'un dispositif technique, d'incorporer celui-ci dans ses gestes et ses postures qui sont liées à ses compétences, son sexe, sa physiologie, etc.) (Marcellini et Miliani, 1999).

Les dispositifs numériques imposent aussi de nouvelles techniques du corps, pour reprendre l'expression de Marcel Mauss, de nouveaux gestes et de nouvelles postures.

> *Ces "habitudes" varient non pas simplement avec les individus et leurs imitations, elles varient surtout avec les sociétés. Il faut y voir les techniques et l'ouvrage de la raison pratique collective et individuelle, là où on ne voit d'ordinaire que l'âme et ses facultés de répétition* (Mauss, 2013, 369).

Les corps des chercheurs doivent s'adapter à une numérisation croissante de leurs pratiques, ils doivent développer une capacité croissante à interagir avec des dispositifs numériques qui ne cessent de se multiplier. La dextérité des mains et des doigts est de plus en plus sollicitée pour naviguer, à travers des écrans tactiles, dans les menus et entre les fonctionnalités d'applications qui permettent d'enregistrer des informations ou d'interagir avec d'autres équipements. Les moyens numériques (ordinateurs, tablettes, smartphones, etc.) imposent donc aux individus des façons de s'en servir, de se mouvoir, ou de rester assis devant un ordinateur, un écran, un clavier, de prendre en main une souris, d'interagir avec un écran par une interface dite « homme-machine », de placer des objets à numériser dans des machines dédiées pour en faire la numérisation en 3D, d'installer des cibles sur les sols de sites archéologiques ou des sphères sur les parois d'art rupestre pour en faire un relevé en photogrammétrie, etc. Ces relations étroites entre les corps des chercheurs, leurs dispositifs matériels et leurs usages a été largement décrite par F. Waquet dans ce qu'elle appelle l'ordre matériel des savoirs qui n'est jamais uniquement matériel (2015).

6. Pour une mise en perspective

Notre projet de recherche consiste à tenter de décoder ce que les appareils, les codes informatiques et les algorithmes font à une expérience de recherche dans une science humaine comme l'archéologie. Entre questionnements épistémologiques, sociologiques, historiographiques et anthropologiques, nous avons tenté de décrire et

d'interroger les effets de la « culture numérique » (Cardon, 2019) sur des pratiques de travail de recherche et sur les identités professionnelles des chercheurs.

En aucune manière, le numérique ne peut se porter candidat au « *solutionnisme technologique* » qu'Evgueny Morozov, chercheur, journaliste et essayiste, spécialiste des implications politiques et sociales de la technologie, a déjà largement critiqué (Morozov, 2014), tout comme l'ont fait de nombreux auteurs en France (Sardin, 2016 ; Biagini, 2012 ; Calan et Cauchard, 2019 ; Filippova, 2019).

Certes, l'usage d'outils informatiques au service des pratiques de la recherche en archéologie fait quotidiennement la démonstration de ses bénéfices. Mais, comme tous les dispositifs techniques, ceux utilisés dans ce domaine ne sont pas neutres. Ils portent en eux leur généalogie, leur origine, etc. Ils ne s'arrogent pas par eux-mêmes leurs propres conditions de possibilité, qui sont fondamentalement liées au contexte socio-techno-culturel des organisations humaines dans lesquelles s'inscrivent ces dispositifs numériques et leurs utilisateurs.

A l'occasion de l'établissement du plan stratégique de l'Inria 2013-2017, le philosophe et historien des sciences Michel Serres, a suggéré que les acteurs de l'informatique soient sensibilisés aux sciences sociales.

Autrement dit, l'informatique produit des réseaux de relations inédites et des institutions à l'état naissant, des individus originaux et des collectifs insolites. […] Non seulement pour l'avenir des recherches propres à Inria, mais aussi pour le futur de nos sociétés, peut-être vaudrait-il mieux que les artisans de l'informatique forment leurs propres chercheurs aux sciences sociales et aux questions éthiques, quitte à les remodeler, plutôt que d'aller chercher dans ces disciplines telles qu'elles existent aujourd'hui, des chercheurs autrement formatés (Serres, 2013)

Cette suggestion du philosophe pourrait conduire à une prise en compte par des acteurs de l'informatique aux différentes dimensions que posent le processus de numérisation du monde qui ne relève pas uniquement d'un projet scientifique ou technologique mais bien anthropologique puisqu'il concerne une nouvelle représentation du monde que ce processus implique.

Avec l'historien et anthropologue François-Xavier Petit, nous pensons que :

Il est temps que la technologie numérique prenne conscience de son pouvoir social (…). Le numérique est matériel. Il a des frontières, une sociologie. Il est temps de le reconnecter avec le sol social (Petit, 2019).

Le risque n'est pas nul, en effet, de voire certaines communautés de chercheurs, notamment en archéologie, faire une confiance aveugle aux machines et à leurs algorithmes. Ceux-ci n'ont de cesse de vouloir rendre plus rapides, plus faciles, plus fluides, plus liquides, les tâches de travail, au nom d'une rationalisation sans fin. Mais quel sens peut encore prendre le travail si celui-ci ne connaît plus aucune limite dans l'accélération du temps dans lequel il prend place ? Comment l'expérience humaine peut-elle se construire de façon supportable si le sens donné au travail se dissout dans des pratiques où les liens de socialisation sont fragilisés ? Jusqu'à quel point une pratique de recherche en science humaine peut-elle s'adosser à des dispositifs techniques, dans une sorte de croyance sans limite dans le pouvoir

des techniques numériques (Gollac et Kramarz, 2000), qui confine parfois à un « fétichisme technologique » (Huggett, 2004) ?

En reprenant le titre de l'ouvrage de J-C. Monod (2019), gageons que pour l'archéologie, comme pour les autres domaines de la recherche française, l'art de ne pas se faire trop gouverner pourrait passer par celui de ne pas se faire trop numériser… ni trop programmer.

Bibliographie

Berry G. (2019) *La pensée informatique*. 2019, Ed. du CNRS, coll. Les Grandes Voix de la Recherche, Paris.

Biagini E. (2012). *L'emprise numérique*. *Comment Internet et es nouvelles technologies ont colonisé nos vies*. Ed. L'Echappée, Paris.

Boullier D. (2019). *Sociologie du numérique.* Armand Colin, Paris.

Bourdieu P. (1979). *La distinction. Critique sociale du jugement.* Les éditions de minuit, Paris.

Buchsenschutz O. (1989). *Expérimentations sur le site du Mont-Beuvray*. Le courrier du CNRS, 73, Paris, p.30.

Calan J. de et Cauchard J. (2019). *Remède contre l'hystérie numérique. Pourquoi la « révolution digitale » n'est pas une révolution*. Robert Laffont, Paris.

Cardon D. (2019). *Culture numérique*. Presses de la Fondation Nationale des Sciences Politiques, Paris.

Casilli A. (2010). *Les Liaisons numériques. Vers une nouvelle sociabilité ?* Seuil, Paris.

Dacheux E. (2018). *L'idéologie numérique contre le sens critique.* ⟨https://theconversation.com/debat-lideologie-numerique-contre-le-sens-critique-94005⟩

Dejours C. (2003). *L'Évaluation du travail à l'épreuve du réel : Critique des fondements de l'évaluation*. INRA, Paris.

Desachy B. (2008). *De la formalisation du traitement des données stratigraphiques en archéologie de terrain.* Thèse en Sciences de l'Homme et Société. Université Panthéon-Sorbonne - Paris I, 2 vol. ⟨http://tel.archives-ouvertes.fr/tel-00406241v2⟩

Desachy B. (2016). *Du carnet de fouilles aux systèmes d'information archéologiques de terrain : quelques remarques sur l'évolution de l'enregistrement et l'impact de l'informatisation.* Colloque « Archivage, publication et mise à disposition de données archéologiques » du consortium MASA, 26 et 27 septembre 2016 à la Nanterre, Maison Archéologie et Ethnologie : https://vimeo.com/189334301

Doueihi M. (2013). *Qu'est-ce que le numérique ?* PUF, Paris.

Dubar C. (1991). *La socialisation, construction des identités sociales et professionnelles*. Armand Colin, Paris.

Dubar C. (2000). *La crise des identités. L'interprétation d'une mutation.* PUF, Paris.

Edgerton D. (2013). *Quoi de neuf ? Du rôle des techniques dans l'histoire globale*. Le Seuil, Paris.

Ellul J. (2012), *Le bluff technologique*. Hachette, Paris.

Eusèbe S. (2019). *Imagerie numérique et représentation des données en archéologie*, http://journals.openedition.org/insitu/21467

Fernandes P. et al. (2013). *Les formations à silex dans le Sud de la France : Élaboration en multipartenariat d'une base de données géoréférencées, premiers résultats*. Dans. Actes de la séance de la Société préhistorique française, Nice, 28-29 mars 2013, Société préhistorique française, pp.137-150, 2016, Séances de la Société préhistorique française. ⟨hal-01436404⟩

Filippova D. (2019). *Technopouvoir. Dépolitiser pour mieux régner*. Les liens qui libèrent, Paris.

Foucault M. (1969). *L'Archéologie du savoir*. Gallimard, Paris.

Foucault M. (2004). *Sécurité, Territoire, Population,* EHESS, Gallimard, Le Seuil, Paris.

Gardin J-C. (1991). *Une contribution des « humanités » à l'informatique : de PENELOPE (1955) à ZETHOS (1974) et au-delà ».* Dans *Le calcul et la raison. Essais sur la formalisation du discours savant.* Ed. de l'EHESS, Paris.

Goffman E. (1973). *La Mise en scène de la vie quotidienne*. Les Editions de Minuit, Paris

Gollac M. et Kramarz F. (2000) *L'informatique comme pratique et comme croyance*. Dans Actes de la recherche en sciences sociales. Vol. 134, septembre 2000. L'informatique au travail. pp. 4-21

Gruel C. et Buchsenschutz O. (1990). *Informatique et archéologie*, Dans Les dossiers d'Archéologie, 153, octobre 1990, pp.80-83

Hey, T., Tansley, S. and Tolle, K. (2009), *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, october 2009, 253 pages

Huggett, J. (2000). *Computers and Archaeological Culture Change* Dans On the Theory and Practice of Archaeological Computing, edited by G. Lock and K. Brown, 5–22. Oxford: Oxford University Committee for Archaeology Monograph 51

Huggett, J. (2004). *Archaeology and the New Technological Fetishism.* Dans Archeologia e Calcolatori 15: 81–92. http://www.archcalc.cnr.it/indice/PDF15/05_Hugget.pdf.

Huggett J. (2020). *Is Big Digital Data Different? Towards a New Archaeological Paradigm.* Dans Journal of Field Archaeology, Volume 45, 2020 - Issue sup1 : https://www.tandfonline.com/doi/full/10.1080/00934690.2020.1713281

Jacob C. (2007 et 2011). *Lieux de savoir.* Tomes 1 et 2. Albin Michel Paris.

Koehler A. et Tufféry C. (2012). *Harmonisation des méthodes et outils pour l'information archéologique à l'Inrap : constats, enjeux et perspectives pour un établissement national*. Dans Archeologia e Calcolatori, All'Insegna del giglio, 2012, pp.229-238. ⟨hal-01853703⟩

Kuhn T. (2008). *La structure des révolutions scientifiques*. Flammarion, Paris.

Le Deuff O. (dir.) (2014). *Le temps des humanités digitales : la mutation des sciences humaines et sociales*. Edition FYP, Paris.

Lerhoëff A. (2019. *L'archéologie*. PUF, Paris.

Marcellini A. et Miliani M. (1999). *Lecture de Goffman*. Dans Corps et culture, Numéro 4 : http://journals.openedition.org/corpsetculture/641

Mauss M. (2013). *Les techniques du corps.* Dans Sociologie et anthropologie, PUF, Paris.

Monod J-C. (2019). *L'art de ne pas être trop gouverné*. Le Seuil, Paris.

Morozov E. (2014) *Pour tout résoudre cliquez ici : L'aberration du solutionnisme technologique*. FYP éditions, Paris.

Mounier P. (dir.) (2012). *Read/Write Book 2 : une introduction aux humanités numériques.* OpenEdition Press, Marseille.

Petit A. (2019). *La recherche, une arme pour les combats du futur,* http://www.lesechos.fr/idees-debats/sciences-prospective/la-recherche-une-arme-pour-les-combats-du-futur-1150759

Petit F-X. (2019). *Il est temps que la technologie numérique prenne conscience de son pouvoir social* https://www.lemonde.fr/idees/article/2019/12/31/il-est-temps-que-la-technologie-numerique-prenne-conscience-de-son-pouvoir-social_6024440_3232.html

Py M. (dir.) (1991). *Système d'enregistrement, de gestion et d'exploitation de la documentation issue des fouilles de Lattes*. Dans Lattara 4, Lattes, 1991, 224 p.

Rosa H. (2010). *Accélération. Une critique sociale du temps*. Ed. La Découverte, Paris.

Rosa H. (2020). *Rendre le monde indisponible*. Ed. La Découverte, Paris.

Rouvroy A. et Berns T. (2009). *Le corps statistique*. P. Daled, Bruxelles.

Sardin E. (2016). *La siliconisation du monde. L'irrésistible expansion du libéralisme économique*. Ed. L'Echappée, Paris

Serres M. (2013). *Vers de nouvelles sciences humaines ?* Dans Inria, Objectif 2020. Plan stratégique 2013, 2017, Paris, p. 4-5

Simondon G. (1989). *Du mode d'existence des objets techniques*. Aubier, Paris.

Tufféry C. et Augry S. (2019). *Harmonisation de l'acquisition des données d'opérations d'archéologie préventive. Retours d'expériences et perspectives à partir de l'application EDArc*. Atelier DAHLIA DigitAL Humanities and cuLtural herITAge : data and knowledge management analys, Jan 2019, Metz, France. ⟨hal-02472817⟩

Tufféry C. (2019). *Les compétences numériques en archéologie : un défi majeur et des risques de déni*, http://www.revue-interrogations.org/Les-competences-numeriques-en

Waquet F. (2015). *L'ordre matériel du savoir. Comment les savants travaillent, XVIe-XXIe siècles*. CNRS Éditions, Paris.

Waquet F. (2019). *Une histoire émotionnelle du savoir. XVIIe-XXIe siècles*. CNRS Éditions, Paris.

Adoption de l'identifiant chercheur ORCID : le cas des universités toulousaines

Marie-Dominique Heusse, Guillaume Cabanac

IRIT UMR 5505 CNRS, Université Paul Sabatier, Toulouse, France
{marie-dominique.heusse,guillaume.cabanac}@irit.fr

RÉSUMÉ. *Les systèmes d'information de la recherche collectent et mettent en visibilité la production scientifique des chercheurs. Leur désambiguïsation est capitale pour ne pas fusionner les productions de plusieurs personnes (cas des homonymes). Or, l'initiative ORCID offre un identifiant à chaque chercheur, pointant vers ses affiliations et sa bibliographie. Les agences de financement (ANR et ERC) et les revues savantes encouragent l'adoption d'ORCID. Nous présentons une méthode pour quantifier cette adoption selon la discipline et de la catégorie d'emploi des publiants d'un établissement. La preuve de concept est réalisée sur les données des 6 471 personnels rattachés aux 150 laboratoires du site toulousain. Nous confrontons avec une validation manuelle leur identité aux 7,3 de millions profils d'orcid.org. Nous observons une adoption croissante d'ORCID avec une disparité d'adoption selon les disciplines. Étonnement, des profils sont uniquement créés pour obtenir un ORCID, sans renseigner ni affiliation ni bibliographie. Ces profils « vides » ont peu d'intérêt pour la tâche de désambiguïsation des identités. À notre connaissance, aucune autre étude de cette ampleur n'a été publiée concernant l'adoption d'ORCID sur un site universitaire multidisciplinaire. La méthode proposée est réplicable et de futures études pourront chercher à confronter les situations et les dynamiques d'évolution.*

ABSTRACT. *Research-focused information systems harvest and promote researchers' scientific output. Disambiguating their identities is key not to merge several persons' records (case of homonyms). ORCID offers an identifier to link one's identity, affiliations, and bibliography. Funding agencies (e.g., ANR, ERC) and scholarly journals promote ORCID. We introduce a method to quantify its adoption according to researchers' discipline and occupation in a higher-education organisation. We semi-automatically matched the 6,471 staff members affiliated to the 150 labs of the Toulouse scientific area with the 7.3 million profiles at orcid.org. The increasing ORCID adoption comes with discipline-wise disparities. Unexpectedly, many profiles are void of information and might have been created only to get an identifier. Those empty profiles are of little interest for the entity disambiguation task. To our knowledge, this is the first study of ORCID adoption at the scale of a multidisciplinary scientific metropole. This method is replicable and future studies can target other cases to contrast the dynamics of ORCID adoption worldwide.*

MOTS-CLÉS : *bibliographie, désambiguïsation, ORCID, identifiant chercheur, multidisciplinarité.*

KEYWORDS: *bibliography, disambiguation, ORCID, researcher identifier, multidisciplinarity.*

1. Introduction

La question des identifiants est au cœur de tout système d'information gérant des individus au sein d'une organisation. Les établissements d'enseignement supérieur et de recherche (ESR) n'échappent pas à la règle, compliquée cependant par la diversité même des activités de leurs membres :

– fonctions administratives, d'enseignement, de recherche ;

– et pour la recherche, multiplicité des acteurs/opérateurs avec lesquels les enseignants-chercheurs et chercheurs sont en interaction : tutelles locales et nationales, agences de financement, éditeurs publics et privés, plateformes d'archives ouvertes, réseaux sociaux académiques, etc.

Dans les faits, deux pratiques coexistent :

– les systèmes de gestion des établissements organisent la communication des informations au sein de modules spécifiques : gestion des ressources humaines (SIRH), gestion des formations, gestion financière, en particulier. Des actions sont menées au niveau national pour améliorer la qualité des données dans ces outils, voir par exemple le projet Sinaps [1] ;

– pour l'activité recherche, le paysage est beaucoup plus éclaté. Si de très rares établissements français commencent à s'intéresser à des outils de type *Current research information system* ou CRIS [2], il n'y a pas pour le moment de réalisations à l'instar des universités d'Europe du Nord ou d'Amérique du Nord. Le repérage des productions de la recherche en particulier s'effectue au travers d'une série de bases de données, revues en ligne, archives ouvertes, qui ont chacune leur propre système d'identification des auteurs, spécifique à la base, comme ResearcherId (Web of Science), AuthorId (Scopus), IdHal (HAL) et IdRef (Sudoc). En parallèle se sont développées des initiatives internationales pour créer des normes d'identifiants univoques et pérennes : ISNI [3] et VIAF [4] notamment.

ORCID, acronyme de *Open Researcher and Contributor ID*, est un système international créé en 2012 et construit sur la norme ISNI dans le but d'identifier de manière unique les auteurs de publications scientifiques et académiques (Haak *et al.*, 2012). Porté par une organisation à but non lucratif – mais à laquelle participent les plus grands éditeurs privés – il a été rapidement adopté par les chercheurs eux-mêmes : au 17 octobre 2019 [5] on dénombre 7 314 172 profils créés tandis que l'UNESCO estime

1. http://www.amue.fr/pilotage/logiciels/sinaps/
2. https://www.eurocris.org/
3. http://www.isni.org/
4. http://viaf.org/
5. Export annuel des profils ORCID en XML sous licence CC0 : https://doi.org/10.23640/07243.9988322.v2, voir aussi https://orcid.org/statistics

à 7,8 millions le nombre de scientifiques (Soete *et al*., 2015, p. 32). Les institutions et les revues ont également adopté ORCID : les agences de financement nationales ou européennes ainsi que nombre d'éditeurs exigent désormais cet identifiant pour les soumissions : dépôts de dossiers de demande de financement, soumission d'articles, revues par les pairs (Hanson *et al*., 2016 ; Dunford, Rosenblum, 2018). Par ailleurs, les universités adhèrent à l'organisation ORCID et mettent en œuvre des actions d'incitation pour leurs chercheurs, dans le contexte de la mise en œuvre d'un CRIS [6] ainsi que pour améliorer leurs dispositifs d'archives ouvertes (Brown *et al*., 2016) ou d'évaluation (Haak *et al*., 2018).

Si les avantages de ce dispositif ne font aucun doute, on peut s'interroger sur le degré d'adoption réel par les chercheurs eux-mêmes. Les freins à l'utilisation ne sont en effet pas négligeables : des alternatives existent en matière de mise en visibilité d'un chercheur, plus connues, populaires, ou (mais pas toujours) plus faciles à mettre en œuvre : profil Google Scholar, réseaux sociaux académiques comme Mendeley, ResearchGate, Academia.edu, voire LinkedIn (Tran, Lyon, 2017 ; French, Fagan, 2019). L'argument d'efficacité du référencement (pérennité garantie de l'identifiant ORCID, assurance de l'élimination des doublons) peut aussi entraîner des réticences chez des universitaires attachés à leur liberté académique et hostiles à des dispositifs de repérage de leur activité. Nous avons donc cherché à mesurer le taux d'adoption d'ORCID dans un cadre circonscrit et connu : les enseignants-chercheurs et chercheurs des établissements constituant l'université de Toulouse, sur la période 2013–2017. Au-delà d'une réponse binaire (adoption : oui ou non), nous avons voulu affiner l'analyse en repérant l'évolution dans le temps, le lien éventuel avec la discipline, ainsi que l'utilisation qui était faite des fonctionnalités d'ORCID : renseignement des données d'affiliation et de biographie, alimentation de la notice avec les publications, ou à l'inverse profil « vide ».

La section 2 présente les principales caractéristiques d'ORCID, et notamment celles qui peuvent influer sur les pratiques des adoptants, au moment de l'inscription et ensuite, pour mettre à jour et alimenter les profils. La section 3 présente notre méthode de collecte, de validation et d'analyse des données relatives aux 6 471 personnels du site toulousain. La section 4 présente la dynamique de l'adoption de l'identifiant ORCID par cette communauté universitaire au regard des catégories d'emploi et des pôles disciplinaires. La section 5 discute les résultats inattendus que nos analyses révèlent, avant la conclusion en section 6.

2. Profil ORCID : création, alimentation et visibilité des profils

La création de profil ORCID peut se faire soit de manière individuelle à l'initiative d'un chercheur ou d'une personne mandatée par lui, soit dans le cadre d'une politique mise en œuvre par une institution qui a contractualisé avec ORCID et bénéficie à ce

6. L'université du Colorado est un des cas emblématiques : https://www.colorado.edu/fis/orcid.

titre d'outils permettant d'automatiser un certain nombre de processus [7]. Fin 2019, ORCID comptait 7,8 millions de profils créés de par le monde, et 1 107 institutions partenaires dont 79 % d'universités et organismes de recherche. Plus de la moitié de ces institutions partenaires sont en Europe, cependant la France n'en compte que six : 2 maisons d'édition, 4 organismes de recherche, et donc aucune université. Un accord de consortium entre ORCID et la France a néanmoins été signé en 2019 dans le but de développer le réseau d'établissements affiliés [8].

Au moment de la création d'un profil ORCID, le chercheur choisit une option de visibilité (figure 1). Le profil sera consultable par quiconque (visibilité publique), ou par des tiers de confiance [9] seulement (visibilité restreinte), ou uniquement par son créateur (visibilité privée). Puis il renseigne les données le concernant dans un certain nombre de champs (figure 2) : *Person identifiers*, *Employment*, *Education and qualification*, *Invited positions and distinctions*, *Membership and services*, *Funding*, et *Works*. De nouveaux champs ont été rajoutés récemment : *Peer-review* et *Research resources*.

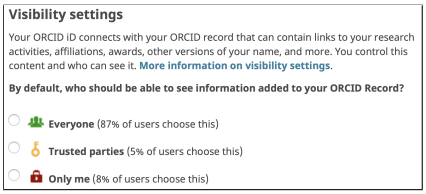


Figure 1. Paramètres de visibilité disponibles lors de la création d'un profil ORCID.

La saisie et notamment la complétude des informations entrées n'étant pas contrôlée, les profils créés peuvent être extrêmement lacunaires. Nous avons rencontré dans les résultats de notre travail un très grand nombre de profils « vides » (section 5.1), sans pouvoir déterminer si c'était parce qu'aucune information n'avait été saisie, ou parce que le chercheur avait fait le choix d'un accès fermé. Cependant, l'information donnée par ORCID (8 % des utilisateurs choisissent la visibilité privée) est bien plus faible que le pourcentage observé à Toulouse, où il semble s'agir surtout de profils non renseignés.

Dans les premières années d'ORCID, la création de doublons (plusieurs profils – donc identifiants – pour un même auteur) était également possible, et nous en avons rencontré dans notre corpus. Cet inconvénient a été repéré et depuis 2017 ce risque de doublons est efficacement détecté au moment de la création d'un nouveau profil [10].

ORCID a assez tôt développé des outils pour faciliter l'alimentation des données sur les publications, et plus récemment sur l'activité d'évaluation par les pairs. Selon le rapport 2019 de l'organisation (Haak *et al.*, 2019, p. 15), près de 67 % des 49 millions de notices de publications qui sont rattachées à des profils l'ont été via une API (Scopus, ResearcherId, Pubmed Central, CrossRef, pour l'essentiel).

7. https://members.orcid.org/api/integrate
8. https://www.couperin.org/services-et-prospective/adhesions-consortiales/orcid
9. À qui le créateur octroie ou révoque ce privilège (cas des plateformes de soumission d'article, notamment).
10. https://orcid.org/blog/2014/01/09/managing-duplicate-iDs

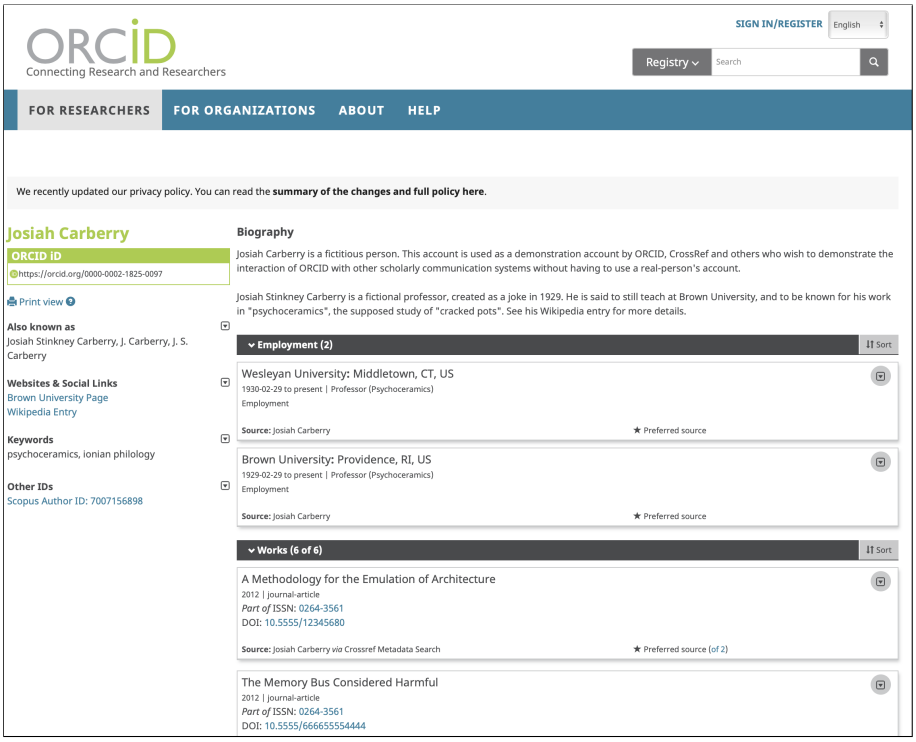


Figure 2. Profil ORCID « modèle » pour le chercheur factice Josiah Carberry.

3. Méthodes et données

Cette section décrit le protocole de collecte des données provenant des établissements ainsi que d'ORCID. Nous détaillons l'appariement automatique de chaque identité de personnel avec les profils ORCID correspondants, puis la démarche de validation manuelle de ces appariements. L'analyse de la base de données originale ainsi constituée permet de révéler la dynamique d'adoption d'ORCID à l'échelle d'une métropole scientifique. C'est à notre connaissance la première étude à cette échelle.

3.1. Collecte des données

Les données sur les 6 471 personnels des laboratoires ont été rassemblées dans le cadre d'un premier travail en 2014–2016 sur la caractérisation de l'activité de recherche du site toulousain (Heusse, 2016). Ce registre des personnels est un tableau comprenant les champs suivants : identité (prénom et nom), établissement, catégorie d'emploi, corps et grade, sexe, année de naissance, laboratoire de rattachement, domaine scientifique. Dans un premier temps, l'ensemble des personnels a été pris en compte, y

compris les personnels BIATSS [11] des catégories A, B et C, car des premiers sondages avaient permis de repérer que certains d'entre eux co-signent des publications et ont un identifiant ORCID. Après rapprochement avec les données ORCID, les personnels de catégorie B et C ont été retirés de la base (130 personnes), à l'exception de ceux identifiés comme publiants, soit 18 personnes. L'ensemble des personnels se répartit en trois grandes catégories d'emploi [12] :

– enseignants-chercheurs ;

– chercheurs : chercheurs des organismes et post-doctorants ;

– autres personnels : ingénieurs d'étude, ingénieurs de recherche, enseignants du second degré, praticiens hospitaliers, notamment.

Les informations, qui sont une agrégation de données fournies par les laboratoires, sont de qualité hétérogène : la date de naissance et le domaine scientifique notamment sont remplis de manière lacunaire, ce qui nous empêche par exemple de repérer pleinement si l'adoption d'ORCID est plutôt le fait de chercheurs jeunes, ou au contraire de ceux qui, montés en responsabilité, l'utiliseraient pour des demandes de financement. Les lacunes sur le domaine scientifique sont compensées par le regroupement des quelques 150 laboratoires du site en six grands pôles disciplinaires, et notre analyse est donc effectuée au niveau de ces ensembles :

– IMI : *Ingénierie, mathématiques et informatique* ;

– SM : *Sciences de la matière* ;

– BSA : *Biologie, santé, agronomie* ;

– SHS : *Sciences humaines et sociales* ;

– DEG : *Droit, économie, gestion* ;

– STU : *Sciences de la Terre et de l'univers*.

La figure 3 montre la diversité des configurations, avec un poids différent de chaque catégorie selon les pôles : les personnels relevant de *Sciences humaines et sociales* et *Droit, économie, gestion* sont caractérisées par la très faible part de chercheurs (organismes et post-doctorants) et de personnels BIATSS en leur sein. Les chercheurs sont à l'inverse légèrement majoritaires en *Biologie, santé, agronomie* et en *Sciences de la Terre et de l'univers*. Enfin, l'importance des autres personnels en BSA s'explique partiellement par le poids de la recherche clinique et le rôle des personnels du CHU dans celle-ci.

Pour chaque identité des 6 471 personnels, nous avons interrogé l'API publique [13] d'ORCID pour obtenir le ou les profils correspondants à partir de la requête composée du prénom et du nom du personnel en question. Certains résultats sont pléthoriques : c'est le cas de la requête « Philippe Durand » restituant 4 534 profils dont 5 profils

11. Personnels ingénieurs, administratifs, techniques, sociaux et de santé et des bibliothèques.

12. Dans la suite de l'article, le terme générique « chercheur » sera utilisé pour désigner l'ensemble des personnels, sauf lorsqu'il s'agit de repérer les situations propres à une catégorie spécifique.

13. https://orcid.org/organizations/integrators/API

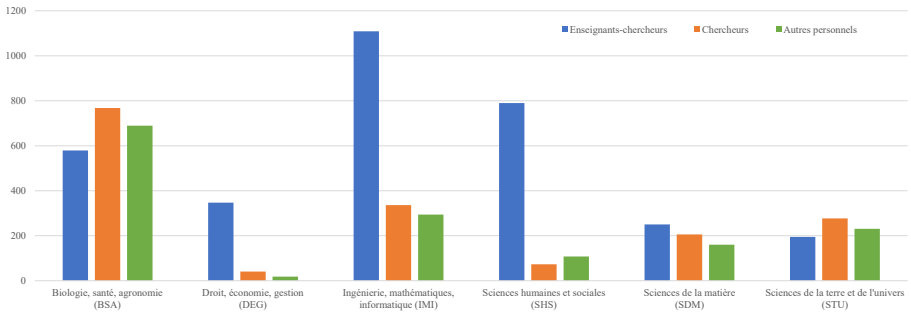


Figure 3. Répartition des 6 471 personnels par pôle et par catégorie d'emploi.

d'homonymes parfaits en tête de liste puis des appariements partiels, tels que « Philippe S. Durand » ou même « Romain Durand ». [14] Nous avons par la suite traité les 20 premiers profils restitués au plus. Chaque profil est un fichier XML [15] contenant, notamment, la date de création du profil, sa date de dernière mise à jour, l'identité déclarée, une biographie, des affiliations et des publications. Le contenu du fichier XML reflète les informations présentées sur le profil consultable en ligne (figure 2).

Nous avons cherché dans les profils une évocation du site toulousain pour désambiguïser les homonymes : ses lieux (Toulouse, Albi, Castres, Rodez et Tarbes) et le nom de ses universités (Capitole, Mirail devenue Jean Jaurès et Paul Sabatier). Cet indicateur était ensuite mobilisé en phase de validation manuelle détaillée dans la section suivante.

3.2. Validation manuelle de la jointure entre registre des personnels et ORCID

Une fois l'interrogation via l'API d'ORCID réalisée, la base de données constituée proposait, pour chaque individu, un des cas de figure suivants :

– aucune correspondance dans la base d'identifiants d'ORCID,

– une seule correspondance trouvée,

– de 2 à 20 correspondances trouvées.

Une validation a été effectuée quand au moins une correspondance a été détectée. La validation était facilitée par l'extraction de données biographiques et d'affiliation lorsqu'elles étaient présentes, ou par le repérage des affiliations sur les articles quand ceux-ci étaient indiqués. La difficulté a été plus grande avec les profils « vides » : une quasi-certitude a pu être établie dans le cas d'un groupe prénom et nom très discriminant, avec au besoin vérification de cohérence de la discipline et de l'absence de deux chercheurs portant le même nom via une interrogation de Google Scholar.

14. Nous avons observé que certaines identités fournies par les laboratoires diffèrent parfois des identités de signature des articles : cas des noms de naissance *versus* noms d'usage, notamment. De fait, l'interrogation par appariement exact (avec guillemets, c.-à-d. "Philippe Durand") n'est pas judicieuse pour ces cas.

15. https://members.orcid.org/api/tutorial/reading-xml

Cette quasi-certitude a été écartée dans le cas de noms très répandus. L'étiquetage de l'appariement a utilisé les codes suivants :

– Pas de numéro ORCID correspondant à l'identité du personnel recherchée,

– `0` : il ne s'agit pas de la même personne (cas des différences entre le groupe prénom et nom recherché *versus* restitué par l'interrogation d'ORCID),

– `1` : le profil ORCID permet de valider l'appariement,

– `?` : profil ORCID vide, mais un seul chercheur correspond après vérification dans Google Scholar,

– `??` : le groupe prénom et nom restitué correspond à l'identité recherchée, mais plusieurs chercheurs correspondent et le profil vide ne permet pas de trancher.

On a donc considéré comme adoptants d'ORCID les personnels étiquetés avec un code `1` ou `?`, soit 2 580 personnes. Il faut noter par conséquent que le corpus retenu représente une fourchette basse de l'adoption d'ORCID : des identifiants ORCID du groupe `??` peuvent correspondre de fait à des chercheurs du site.

4. Résultats quantitatifs

Le taux d'adoption d'ORCID pour l'ensemble de la population concernée, mesuré début mars 2020, est de 39,87 %. En excluant les « autres personnels », ce taux est de 44,6 %. Ces deux taux recouvrent des disparités importantes que nous avons souhaité mettre en lumière dans les sections suivantes.

4.1. Adoption de l'identifiant ORCID par catégorie d'emploi

Comme il apparaît dans la figure 4, les chercheurs (non enseignants) sont sans grande surprise les adoptants majoritaires (à 55 %). Le taux plus faible des enseignants-chercheurs (39,2 %) s'explique sans doute en partie par leur surreprésentation dans les disciplines de *Sciences humaines et sociales* et de *Droit, économie, gestion* où les exigences des revues et des tutelles vis-à-vis d'ORCID sont plus faibles. Pour les « autres personnels » (24,2 %), s'ils prennent part à la recherche menée dans les équipes, peu d'entre eux sont associés aux publications. Rappelons qu'il s'agit ici, pour les BIATSS, des catégorie A et des 18 publiants de catégorie B et de catégorie C, ainsi que des praticiens hospitaliers et des professeurs du second degré, principalement.

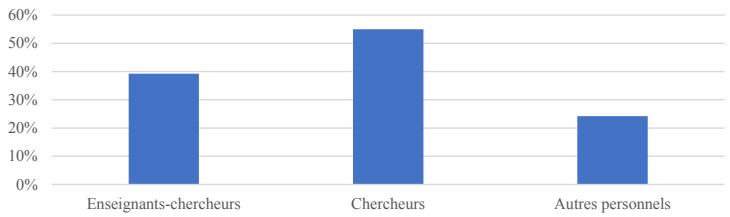


Figure 4. Pourcentage d'adoption d'ORCID par catégorie d'emploi.

4.2. Adoption de l'identifiant ORCID par pôle disciplinaire

Le taux d'adoption par pôle disciplinaire (figure 5) est en cohérence avec ce qu'on sait des pratiques de publication, des politiques des revues dans les différents grands champs disciplinaires, et de l'usage des identifiants plus ancien et plus fort en sciences exactes. L'analyse est à compléter cependant par la figure 8, qui représente les courbes respectives d'adoption dans le temps pour chacun des pôles.

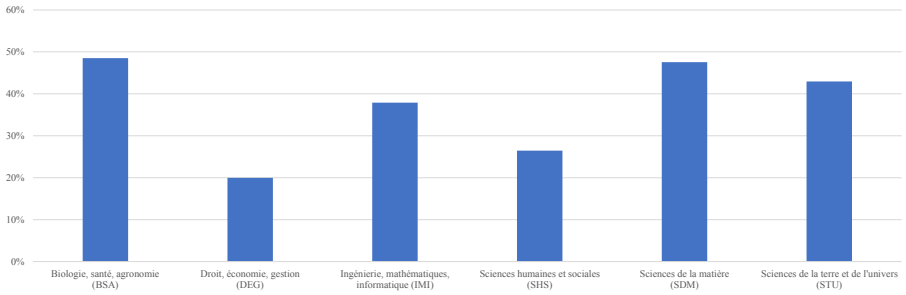


Figure 5. Pourcentage d'adoption d'ORCID par pôle disciplinaire.

La figure 6 combine les taux d'adoption par pôle et par catégorie d'emploi. Elle permet de constater que le taux d'adoption plus élevé pour les chercheurs se retrouve de façon assez homogène dans les différents pôles disciplinaires, même s'il est plus faible que la moyenne en *Droit, économie, gestion* : il faut sans doute y voir l'impact des consignes des organismes de recherche en la matière [16]. Pour les enseignants-chercheurs, elle montre nettement le plus faible pourcentage en SHS et DEG, déjà évoqué plus haut. Les autres personnels ont des taux significativement plus élevés en BSA et SDM, et dans une moindre mesure en STU : dans ces pôles, plusieurs laboratoires ont été repérés comme associant des personnels BIATSS comme cosignataires des articles de recherche (section 5).

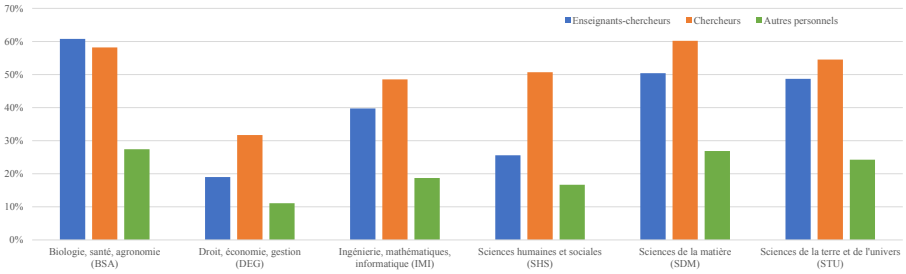


Figure 6. Pourcentage d'adoption d'ORCID par pôle disciplinaire et par catégorie d'emploi.

16. Fiche DORANUM https://doranum.fr/wp-content/uploads/Fiche-ORCID.pdf

4.3. Adoption progressive de l'identifiant ORCID : analyse longitudinale

4.3.1. Évolution globale

La figure 7 montre que loin d'être linéaires, les créations s'articulent sur le calendrier universitaire avec les échéances de dépôts de projets ANR [17], et que cette tendance paraît se renforcer avec les années, les mois d'octobre constituant des pics de création de manière de plus en plus visible. Cette tendance est probablement à rapprocher de la pratique de création de profils « vides », la motivation du chercheur étant principalement d'obtenir un identifiant pour compléter le dossier de demande de financement.

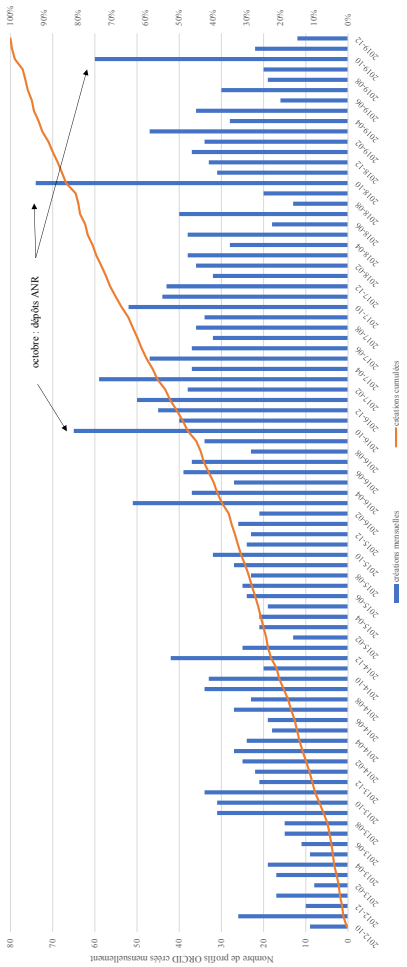


Figure 7. Créations de profils ORCID au fil du temps.

17. cf. page 20 du guide : https://anr.fr/fileadmin/aap/2019/aapg-anr-2019-Guide.pdf

4.3.2. Évolution par pôle disciplinaire

La figure 8 représente la courbe d'évolution des créations de profils ORCID pour chaque pôle disciplinaire. Les formes respectives des courbes sont dans l'ensemble homogènes, mais permettent cependant de repérer certaines caractéristiques :

– les disciplines STU et BSA ont été au début (et sur presque toute la période pour STU) les plus actives dans les créations,

– d'autres, comme SDM, sont parties moins vite, mais ont connu une forte accélération dans les dernières années. C'est aussi le cas de SHS, marqué par des taux faibles au départ, mais qui a fini par combler ce décalage.

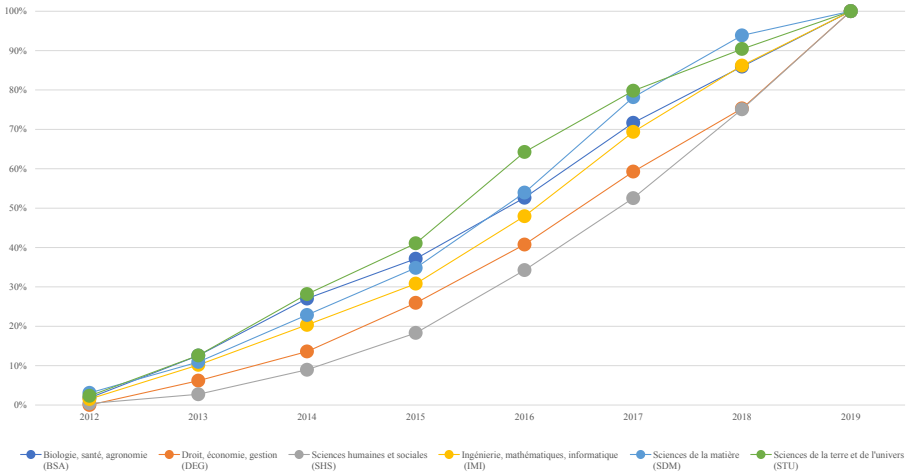


Figure 8. Part cumulée d'adoption d'ORCID par pôle disciplinaire. La part des profils correspondant à $y = 100\,\%$, pour un pôle disciplinaire donné, est présentée en figure 5.

5. Observations qualitatives

Cette section complète l'étude quantitative des données par des observations qualitatives liées à la diversité des usages des profils ORCID. Nous commentons également quelques cas de mésusages notables puis réalisons un focus sur la catégorie d'emploi des « autres personnels ».

5.1. Diversité des (més)usages des profils ORCID

On constate une très grande variété dans les modes d'appropriation et d'usage d'ORCID. Rappelons que l'inscription se fait à l'initiative du chercheur, et qu'il a la possibilité de renseigner les champs énumérés à la section 2. Des profils ORCID détaillés comprenant à la fois des données biographiques et de publications ne sont cependant pas la majorité : certains indiquent seulement les éléments d'affiliation,

d'autres uniquement les publications. Surtout, nombre de profils sont complètement vides [18] (figure 9).

Figure 9. Exemple d'un profil ORCID « vide ».

Comme indiqué précédemment, les profils vides ont deux origines : profil non renseigné ou profil dont la visibilité est privée. Répartir les profils de notre corpus dans ces deux catégories n'est pas possible techniquement. Indiquons seulement que la fréquence des profils vides dans notre corpus s'élève à 40 %, alors qu'ORCID indique qu'il y a 8 % de profils privés (figure 1). Il serait intéressant de repérer la variété des pratiques, du profil complet au profil vide, et les priorités éventuellement données par les chercheurs aux données d'affiliation ou de publications. À titre d'exemple, la figure 10 présente le pourcentage de profils ORCID listant au moins une publication, par pôle disciplinaire.

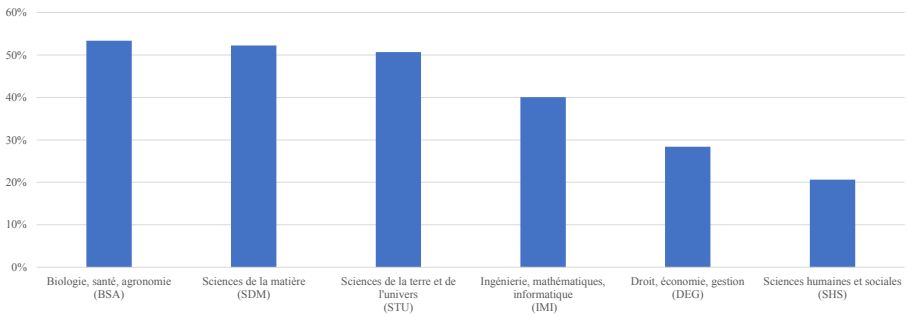


Figure 10. Part des profils ORCID listant au moins une publication.

18. Il y a, par exemple, cinq profils distincts pour l'identité « Philippe Durand » dans ORCID, tous vides…

Nous avons aussi observé des pratiques étonnantes, aussi bien en matière d'affiliation que de signalement des publications, sans toutefois que ces pratiques aient été quantifiées :

– les sections *Biography* et *Employment* du profil, qui permettent de repérer les mobilités dans la carrière d'un chercheur et donc l'évolution de ses affiliations, sont parfois détournées pour mentionner les universités étrangères avec lesquelles il a des liens d'échange mais qui ne sont pas ses « employeurs » à proprement parler. Il arrive même que la rubrique énumère seulement ces différents établissements prestigieux (cas des universités américaines, notamment) mais sans indiquer celui de rattachement (c'est-à-dire l'employeur). Pour les chercheurs rattachés à des organismes nationaux, d'autre part, il n'est pas rare de trouver « CNRS Paris » ou « INSERM Paris », sans aucune référence à leur implantation régionale ni à leur laboratoire sur le site – que l'on trouve seulement en consultant les articles, lorsqu'ils sont présents.

– le nombre de publications que les chercheurs ajoutent à leur profil ORCID sous la section *Works* varie d'une seule à plusieurs centaines. À l'évidence, certains ne maîtrisent pas les processus d'alimentation automatique de cette section, et la qualité des données en pâtit (pas de date de publication, pas de DOI…). Nous avons réalisé des sondages pour mesurer l'écart entre la production réelle d'un auteur et le nombre d'articles signalés dans ORCID, et cet écart peut atteindre, lui aussi, plusieurs centaines d'articles.

La création de plusieurs profils ORCID par la même personne, non contrôlée jusqu'en 2017 [19], a été également repérée. Nous avons quantifié cette pratique, qui reste cependant très minoritaire : il y a quinze profils ORCID correspondant à des doublons certains dans notre corpus : 8 cas de doublons entre en un profil renseigné et un profil vide, 7 cas entre deux profils vides. Plus surprenante est la découverte de 3 autres doublons où les deux profils sont à chaque fois renseignés pour la même personne – avec une quantité de données inégale, cependant.

Une hypothèse d'interprétation de la diversité de ces pratiques serait la « concurrence » entre les différents dispositifs de mise en visibilité de la production des chercheurs, et le fait pour eux d'en être plus familiers ou plus convaincus de leur efficacité. Pour explorer cette hypothèse sur une échelle très réduite, on a mis en regard utilisation d'ORCID *versus* création d'un profil Google Scholar pour 15 chercheurs du site distingués au cours des années récentes comme *Highly Cited Researchers* par Clarivate Analytics [20] (tableau 1).

On le voit : aucune corrélation n'apparaît avec certitude, même pour des chercheurs familiers des consécrations comme ceux du palmarès de Clarivate Analytics.

19. Depuis, le contrôle est réalisée sur l'adresse email, voir note 10.

20. https://recognition.webofsciencegroup.com/awards/highly-cited/2019/ et années antérieures.

Tableau 1. Présence en ligne sur ORCID et Google Scholar de 15 chercheurs listés comme Highly Cited Researchers *(HCR) par Clarivate Analytics. La complétude des profils ORCID est indiquée : existence d'un profil, renseignement de la bibliographie, de l'affiliation et des publications.*

n°	discipline	ORCID				Google Scholar
		profil	biographie	affiliation	publications	
1	agronomie	×		×	×	×
2	biologie	×		×		
3	écologie					×
4	écologie	×		×		×
5	économie					×
6	ingénierie	×		×	×	×
7	santé	×				×
8	santé	×			×	×
9	santé	×	×		×	
10	santé					
11	santé	×				
12	santé	×				
13	santé	×		×	×	
14	santé	×			×	
15	terre et espace	×				×

5.2. Focus sur les créations de profils ORCID par les « autres personnels »

Moins de 25 % des individus de la catégorie d'emploi « autres personnels » a créé un ORCID au cours de la période (section 4.1). Ce taux global recouvre des disparités, et celles-ci sont très clairement liées aux pratiques de publication, et d'association – ou pas – de l'ensemble des personnels ayant pris part à la recherche en qualité de cosignataires de l'article. Nos données permettent de repérer les disciplines qui semblent considérer cette association comme normale : en tête vient le pôle *Biologie, santé, agronomie* où cette pratique se rencontre dans 7 laboratoires sur les 30 qui constituent le pôle, et représente visiblement une caractéristique d'un certain nombre de recherches dans le domaine de la santé ; on la trouve ensuite dans les différents laboratoires du pôle *Sciences de la Terre et de l'univers*, dans plusieurs laboratoires du pôle *Sciences de la matière* (chimie, matériaux…) et enfin dans le pôle *Ingénierie, mathématiques et informatique* (robotique, mécanique des fluides). La création d'un ORCID par ces

personnels est un corollaire logique de la reconnaissance qui leur est accordée par les autres chercheurs du laboratoire.

6. Conclusion et perspectives

L'identifiant ORCID introduit en 2012 pour aider à la désambiguïsation des identités de chercheurs s'est progressivement inscrit dans le paysage de l'enseignement supérieur et de la recherche. Cet article est la première étude de l'adoption d'ORCID sur le périmètre d'une métropole scientifique majeure à l'échelle nationale. Analysant les profils des 6 471 personnels du site toulousain, notre étude montre une dynamique d'adoption croissante quelle que soit la discipline des chercheurs. À l'heure actuelle près de 40 % de la population étudiée possède un profil ORCID. L'analyse par pôle disciplinaire montre des disparités d'adoption : entre 60 % des enseignants-chercheurs en *Biologie, santé, agronomie* et 20 % des enseignants-chercheurs en *Droit, économie et gestion*. Plusieurs raisons peuvent expliquer ces différences, dont les recommandations voire injonctions inégales des établissements, des sociétés savantes, des agences de financement et le degré d'internationalisation des champs de recherche. Une analyse qualitative de l'adoption d'ORCID, pour les 60 % des profils renseignés, montre qu'ils sont partiellement remplis : les sections disponibles pour décrire la biographie, la trajectoire institutionnelle, les autres identifiants, les financements obtenus… sont très souvent vides et il manque bien souvent de nombreuses publications. Par ailleurs, nous avons observé des mésusages dans la façon d'utiliser et de renseigner les différents champs de la notice ORCID, l'une des omissions les plus critiques étant l'information sur *Employment*, c'est-à-dire l'affiliation du chercheur. Ces observations questionnent quant à la perception de la finalité de l'identifiant ORCID par une partie de la communauté savante.

Il est envisagé de mener des travaux plus qualitatifs pour compléter ce premier travail : explorer via des enquêtes les pratiques et les comportements des usagers, estimer le taux de complétude des publications signalées dans ORCID en comparant avec le Web of Science par exemple ou avec le CV exhaustif fourni par le chercheur lui-même (Youtie *et al*., 2017), repérer l'impact des stratégies d'établissement (par ex., dépôt des publications de l'INRA dans son archive ouverte institutionnelle ProdINRA [21]).

Étant donnée la politique actuelle menée par les opérateurs de l'enseignement supérieur et de la recherche visant à favoriser l'adoption généralisée d'ORCID au sein des établissements, cette étude permet d'estimer l'état actuel de l'adoption tout en soulignant les écueils rencontrés à ce jour, et permet donc de dégager des orientations pour les actions à conduire.

21. http://wiki.inra.fr/wiki/prodinra/Qualite/ORCID+ResearcherId

Bibliographie

Brown J., Demeranville T., Meadows A. (2016). Open access in context: Connecting authors, publications and workflows using ORCID identifiers. *Publications*, vol. 4, n° 4, p. 30. doi:10.3390/publications4040030

Dunford R., Rosenblum B. (2018). Keeping it authentic: Reconciling ORCID iDs gathered at submission with the author manuscript. *Learned Publishing*, vol. 31, n° 3, p. 236–240. doi:10.1002/leap.1159

French R. B., Fagan J. C. (2019). The visibility of authority records, researcher identifiers, academic social networking profiles, and related faculty publications in search engine results. *Journal of Web Librarianship*, vol. 13, n° 2, p. 156–197. doi:10.1080/19322909.2019.1591324

Haak L. L., Fenner M., Paglione L., Pentz E., Ratner H. (2012). ORCID: A system to uniquely identify researchers. *Learned Publishing*, vol. 25, n° 4, p. 259–264. doi:10.1087/20120404

Haak L. L., Meadows A., Brown J. (2018). Using ORCID, DOI, and other open identifiers in research evaluation. *Frontiers in Research Metrics and Analytics*, vol. 3. doi:10.3389/frma.2018.00028

Haak L. L., Petro J. A., Simpson W., Demeranville T., Wijnbergen I., Hershberger S. *et al.* (2019). *ORCID 2019 annual report*. Rapport technique. ORCID. Consulté sur https://doi.org/10.23640/07243.12009153.v1

Hanson B., Lawrence R., Meadows A., Paglione L. (2016). Early adopters of ORCID functionality enabling recognition of peer review: Two brief case studies. *Learned Publishing*, vol. 29, n° 1, p. 60–63. doi:10.1002/leap.1004

Heusse M.-D. (2016). Faire parler les données sur la recherche grâce au web sémantique : le projet VIVO. In *Actes de l'atelier Valorisation et Analyse des Données de la Recherche (VADOR) organisé dans le cadre du congrès INFORSID*, p. 19–25.

Soete L., Schneegans S., Eröcal D., Angathevar B., Rasiah R. (2015). A world in search of an effective growth strategy. In S. Schneegans (Ed.), *UNESCO Science Report: Towards 2030*, p. 20–55. Paris. Consulté sur http://unesdoc.unesco.org/images/0023/002354/235406e.pdf

Tran C. Y., Lyon J. A. (2017). Faculty use of author identifiers and researcher networking tools. *College & Research Libraries*, vol. 78, n° 2, p. 171–182. doi:10.5860/crl.78.2.171

Youtie J., Carley S., Porter A. L., Shapira P. (2017). Tracking researchers and their outputs: New insights from ORCIDs. *Scientometrics*, vol. 113, n° 1, p. 437–453. doi:10.1007/s11192-017-2473-0

Identification de clés pour le succès de projets de gestion informatisée de données environnementales à partir du logiciel Collec-Science

Eric Quinton [1], **Christine Plumejeaud-Perreau** [2], **Sylvie Damy** [3]

1. *INRAE - Unité de recherche Écosystèmes aquatiques et changements globaux*
   50, avenue de Verdun
   33612 CESTAS, France
   eric.quinton@inrae.fr

2. *Littoral Environnement et Sociétés, U.M.R. 7266*
   2 rue Olympe de Gouges
   17000 La Rochelle, France
   christine.plumejeaud-perreau@univ-lr.fr

3. *Université de Bourgogne Franche-Comté – U.M.R. 6249 – Laboratoire Chrono-environnement*
   16 route de Gray
   25030 Besançon cedex, France
   sylvie.damy@univ-fcomte.fr

RÉSUMÉ. *Dans le cadre d'une recherche qui s'ouvre aux pratiques de l'open-data pour la diffusion de données de qualité, il devient plus que jamais nécessaire de doter le monde académique des « bons » outils pour cet objectif ambitieux. Nous parlons de logiciels dédiés à la gestion qui peuvent être soit créés dans le laboratoire où ils sont utilisés, soit diffusés vers une communauté de chercheurs, ou encore pris « sur étagère » dans le catalogue des logiciels libres. A travers cette contribution qui s'appuie sur la grille d'analyse Business Model Canvas, nous tentons d'identifier les points clés de la réussite de tels projets. Nous soulignons notamment la nécessité d'une estimation réaliste des moyens requis pour la diffusion et l'appropriation de tels logiciels et des différents profils de personnels nécessaires à la réalisation de ces projets, avec un focus particulier sur le rôle de curateur / administrateur / animateur des données dans les laboratoires, la nécessité de proposer une offre d'hébergement et la structuration de l'appui.*

ABSTRACT. *In the context of research that is opening up to open-data practices for the dissemination of quality data, it becomes more necessary than ever to equip the academic world with the right tools for this ambitious objective. We are talking about software dedicated to management that can either be created in the laboratory where it is used, or distributed to a community of*

researchers, or taken off the shelf in the open-source software catalogue. Through this contribution, which is based on the Business Model Canvas analysis grid, we try to identify the key points for the success of such projects. In particular, we underline the need for a realistic estimation of the means required for the dissemination and appropriation of such software and the different profiles of personnel needed to carry out these projects, with a particular focus on the role of curator / administrator / data manager in the laboratories, the need to propose a hosting offer and the structuring of support.

MOTS-CLÉS : *Génie Logiciel; Cycle de vie de la donnée - du logiciel ; Administrateur des données*

KEYWORDS: *Software Engineering; Data Life Cycle - Software; Database administrator*

1. Introduction

En juillet 2018, la Ministre de l'Enseignement Supérieur, de la Recherche et de l'Innovation a présenté le plan national pour la science ouverte [1]. L'idée principale de ce plan est la diffusion sans entrave des publications et données issues de recherches financées sur fonds publics. Il propose de structurer et d'ouvrir les données de la recherche, en rendant obligatoire la diffusion ouverte de celles qui sont issues de programmes financés par appels à projets sur fonds publics, en créant la fonction d'administrateur des données et le réseau associé au sein des établissements. Ce plan crée également les conditions pour promouvoir l'adoption d'une politique de données ouvertes associées aux articles publiés par les chercheurs. Actuellement, si certaines disciplines telles que l'astronomie ont déjà une expérience reconnue dans la gestion des données, dans le domaine environnemental où les données ne sont pas forcément très volumineuses mais très hétérogènes (*The long tail data* [2]), la gestion des données reste bien souvent à la charge des structures de recherche telles que les laboratoires.

Les données environnementales sont produites dans un but précis mais peuvent être utilisées pour un autre. La problématique de la qualité, de l'interopérabilité et de la traçabilité de celles-ci est ainsi très forte. La production de logiciels pour les gérer est essentielle afin d'aider les chercheurs à maîtriser au mieux leurs données. Ils sont à forte valeur ajoutée, mais avec un public potentiellement restreint et très spécialisé. Les laboratoires cherchent à embaucher des développeurs pour concevoir de nouveaux outils, et de nombreuses offres sont proposées dans les réseaux spécialisés universitaires. Cela montre l'importance de ce besoin mais aussi le manque de ressources humaines dans ce domaine.

1. https://www.ouvrirlascience.fr/plan-national-pour-la-science-ouverte/
PLAN NATIONAL POUR LA SCIENCE OUVERTE

2. https://www.rd-alliance.org/groups/long-tail-research-data-ig/wiki/plenary-3-long-tail-research-data.html - RDA : Plenary 3 - the long tail of research data

Une alternative à cette création est l'utilisation de logiciels créés par d'autres structures. L'État français, en encourageant la publication des codes en *Open-Source* et en portant à la connaissance du public les différentes réalisations, permet à de nombreux laboratoires de s'appuyer sur des produits tiers à moindre coût. Toutefois, il est illusoire de penser que la simple installation d'un logiciel soit suffisante pour garantir son usage sur le long terme. Mendoza *et al*. (2010) ont identifié trois phases essentielles pour utiliser un logiciel dans ce contexte. La première consiste en une phase exploratoire, pour identifier les produits susceptibles de répondre au besoin. La seconde vise à le tester, voire à l'adapter. Ce n'est qu'à la fin de ces étapes qu'il pourra être utilisé en conditions réelles (*cf.* figure 1). Enfin, une fois que l'appropriation est terminée, le logiciel est utilisé en « routine ».

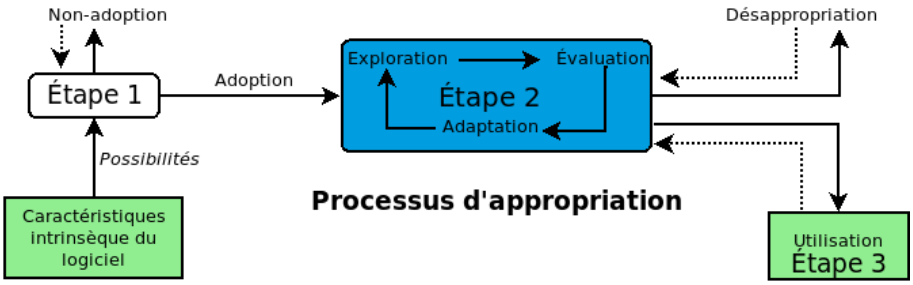


Figure 1. Processus d'appropriation des logiciels (d'après Mendoza et al. *(2010)).*

À chaque étape, différents freins peuvent amener à abandonner soit le choix du logiciel initial, soit le projet lui-même. Les raisons évoquées ont été étudiées : participation insuffisante de l'utilisateur final, mauvaise spécification des besoins, mauvaise estimation des coûts, relations entre les intervenants, fonctionnement de l'équipe de projet, politique de l'entreprise en contradiction avec le projet, etc. (Eberendu, 2015).

Le logiciel Collec-science [3] qui a été créé pour gérer les échantillons prélevés dans le cadre des campagnes scientifiques, a, dès le début du projet, été conçu pour être diffusé auprès d'autres laboratoires. Ainsi de nombreux travaux préparatoires ont été menés en partenariat avec le réseau des Zones Ateliers [4], notamment pour le choix de certaines technologies (choix des types de codes-barres, des matériels, identification des besoins complémentaires, etc.)(Quinton *et al*., 2018). Il a été déployé notamment dans le cadre des Zones-Ateliers. De nombreuses actions ont été nécessaires pour faciliter son déploiement et son appropriation par les différentes unités de recherche intéressées. Un groupe de travail a été constitué et a été largement supervisé par une personne du groupe (Plumejeaud-Perreau *et al*., 2019). Fin 2019, on peut estimer à un vingtaine le nombre de laboratoires de recherche qui soit l'utilisent en routine (étape 3, *cf.* figure 1), soit l'évaluent (étape 2). Cette diffusion n'aurait pas pu voir le jour

3. https://www.collec-science.org

4. Réseau interdisciplinaire de laboratoires français sur l'environnement et les socio-écosystèmes https://inee.cnrs.fr/fr/zones-ateliers

sans la mobilisation d'un certain nombre d'acteurs et la réalisation d'actions (formation, création d'un site web, etc.). Celles-ci ont été mises en place au fur et à mesure des besoins ressentis par l'équipe de diffusion et en fonction des retours des différents utilisateurs potentiels. Leur identification devrait permettre de mieux cerner les conditions qui facilitent l'appropriation d'un logiciel dans un contexte de laboratoire de recherche.

Pour identifier les facteurs de réussite de la diffusion du logiciel Collec-Science, nous avons cherché à répondre à la question suivante : quels sont les acteurs à mobiliser et les tâches à réaliser ? Pour cela, nous avons utilisé un outil largement répandu dans le monde des affaires, le *Business Model Plan* (Osterwalder, Pigneur, 2011). Basé sur un schéma composé de neuf cases, il s'adapte bien à l'étude que nous voulions mener. Nous expliciterons ce choix dans la partie *méthodologie*. Trois cas d'utilisation ont été étudiés : d'une part, la réalisation du logiciel pour un usage interne, d'autre part, sa mise en place dans un laboratoire qui ne l'a pas conçu, et enfin, sa diffusion à d'autres partenaires.

2. Méthodologie

2.1. Business Model Canvas

Pour identifier le processus de conception et de mise en œuvre d'un logiciel, nous avons choisi de nous appuyer sur le *Business Model Canvas*, créé par Alexander Osterwalder (Osterwalder, 2004). Ce modèle est plutôt dédié au monde économique. Cette modélisation des processus n'est pas la première proposée, mais son approche synthétique, basée sur un tableau composé de 9 cases, comme le montre la figure 2, permet d'appréhender et de représenter sur une seule page toutes les tâches, les coûts et les bénéfices attendus d'un projet. Il a d'ailleurs été utilisé dans le cadre de développements open-source de produits matériels (Fjeldsted *et al*., 2012). Luoma *et al*. (2012), dans une étude sur le modèle économique des plates-formes de type SAAS (*Software as a service*), considère que ce modèle est un des plus utilisés pour analyser les sociétés en technologie de l'information. Sa flexibilité nous a permis de l'adapter facilement au contexte de notre étude. Il est d'autre part assez simple d'utilisation, très synthétique et surtout nous permet de bien montrer aux directions des structures de recherche ce qu'il nous faut en termes de ressources humaines pour mener à bien de tels projets. Le tableau 1 récapitule l'ensemble des rubriques du canevas et la signification que nous lui avons donné.

Dans la pratique, le tableau est construit en positionnant des Post-it© sur un canevas affiché au mur , qui est retranscrit ensuite dans un format numérique.

2.2. Les études de cas

La démarche pour mettre en œuvre un logiciel n'est pas tout à fait la même selon qu'il est créé en interne, ou pris « sur étagère ». Dans le premier cas, il faut disposer

Tableau 1. Liste des rubriques disponibles dans un Business Model Canvas et signification dans le contexte de l'étude

Rubrique	Description
Offre	Produit créé et production de valeur attendue
Partenaires clés	Principaux partenaires mobilisés
Activités clés	Principales activités à mettre en œuvre pour produire l'offre
Ressources clés	Principales ressources (humaines ou non) à mobiliser pour réaliser les activités clés
Segments de clientèle	Cible à qui est destinée l'offre
Canaux de distribution	Moyens à mettre en place pour faire connaître l'offre
Relation client	Mécanismes permettant au segment de clientèle de faire remonter ses besoins ou de lui apporter les compléments nécessaires (formation, par exemple)
Structure des coûts	Coûts financiers ou humains à mobiliser pour arriver à l'objectif
Sources de revenus	Gains soit financiers, soit en efficacité, soit en réputation atteints par la diffusion de l'offre

des ressources internes pour la définition des besoins et le codage, mais cela permet d'obtenir une solution proche de la demande initiale. De plus, la formation des utilisateurs est aisée, le développeur étant disponible pour répondre aux questions. Dans le second cas, si l'étape du développement est ignorée, d'autres tâches apparaissent. Il faut s'approprier la structure et les buts du logiciel, le déployer, le configurer, se former et former les utilisateurs finaux, etc. Les ressources à mobiliser sont différentes mais toutes aussi indispensables pour la réussite du projet. Enfin, quand un laboratoire décide de proposer son logiciel à la communauté, il ne suffit pas de le publier pour qu'il soit utilisé. La protection du code, le choix de la licence de diffusion n'en sont que la première étape. Il faut ensuite rédiger des documentations, voire créer des sites vitrines, avoir une stratégie de communication, faciliter le déploiement, etc. Pour ces trois cas, connaître l'ensemble des acteurs à mobiliser et des tâches à exécuter devient nécessaire, pour d'une part faciliter la prise de décision initiale, et d'autre part dimensionner correctement les moyens à allouer.

Nous proposons l'étude de la conception et de la mise en place d'un logiciel de gestion de données dans trois cas différents, à l'aide des *Business Model Canvas*. La première modélisation traitera du développement et de l'utilisation du logiciel *Otolithe* au sein du laboratoire EABX[5]. La seconde s'attachera à identifier les acteurs et les tâches impliqués dans l'installation et l'utilisation du logiciel *Collec-Science* dans le laboratoire de Chrono-Environnement. Enfin, la dernière présentera l'ensemble des

5. laboratoire Écosystèmes Aquatiques et Changements Globaux – INRAE – CESTAS (Gironde)

tâches et des acteurs impliqués à la fois dans la conception et la diffusion du logiciel *Collec-Science*.

Le logiciel Collec-Science a, dès le début de sa conception, été prévu pour être diffusé dans d'autres laboratoires : s'appuyer sur celui-ci pour présenter les tâches correspondant au premier cas d'utilisation aurait soit été théorique, soit aurait présenté un biais (contacts nombreux avec d'autres laboratoires potentiellement intéressés).

3. Modélisation des cas d'utilisation

3.1. Création du logiciel et utilisation en interne dans le laboratoire

Le *Business Model Canvas* présenté en figure 2 correspond au logiciel Otolithe, un outil développé et déployé au sein du laboratoire EABX pour faciliter la lecture des stries de croissance sur les pièces calcifiées de poissons (otolithes, écailles, sections de rayons de nageoire). Les deux logiciels, utilisés comme supports dans cet article, sont très proches à la fois sur le plan technique (mêmes langages de programmation, mêmes outils, mêmes moteurs de bases de données) et sur le plan de la réalisation : ils ont été écrits par le même développeur, dans le même laboratoire. On peut considérer que les résultats obtenus pour Otolithe sont assez équivalents à ceux qui auraient été acquis dans le cas où Collec-Science n'aurait pas été destiné à être diffusé.

Business Model Canvas – Logiciel Otolithe

<table>
<tr>
<td rowspan="2">Partenaires clés</td>
<td>Activités clés
Fourniture des moyens (A)
Développement du logiciel Otolithe (B)
Administration de la plate-forme de production, sauvegarde (C)</td>
<td rowspan="2">Offre (proposition de valeur)
Lecture des pièces calcifiées :
gagner du temps
Fiabiliser la lecture
gagner en traçabilité</td>
<td>Relation client
Formation et appui vers les administrateurs métiers (B)
Formation des utilisateurs (E)
Assistance aux utilisateurs (E)
Gestion de tickets pour les demandes d'évolution et les bugs (B, E)</td>
<td rowspan="2">Segments de clientèle
(D) : Scientifiques et techniciens procédant aux lectures des pièces
(E) : techniciens et scientifiques préparant les photos et procédant à la saisie des données nécessaires à la lecture</td>
</tr>
<tr>
<td>Ressources clés
(A) : E. Rochard – Directeur de l'UR – Chef de projet
(B) : E. Quinton – Admin BDD – développeur
(C) :J. Foury – Admin système</td>
<td>Canaux de distribution
Communication interne (A)</td>
</tr>
<tr>
<td colspan="2">Structure des coûts
Coûts de main d'œuvre</td>
<td colspan="3">Sources de revenus
Gain de temps dans les activités de recherche
Fiabilisation des données récoltées</td>
</tr>
</table>

Figure 2. Business Model Canvas de l'application Otolithe.

Le développement a été mis en place sous l'égide du directeur d'unité. Une fois l'application prête, elle a été mise en production par l'administrateur système. L'appli-

cation a été présentée aux utilisateurs lors de réunions, des formations ont été organisées pour les aider à la prendre en main. Un système de tickets, associé à la plate-forme de gestion de code, leur a été ouvert pour qu'ils puissent faire part des bogues rencontrés ou des évolutions qu'ils souhaitaient. Les gains de productivité pour les équipes techniques et pour les scientifiques chargés de la lecture des pièces calcifiées, la fiabilisation des données (stockage maîtrisé notamment) ont largement surpassé le temps d'écriture du logiciel.

3.2. Mise en place d'un logiciel conçu par un autre laboratoire

Le *Business Model Canvas* (figure 3) a été élaboré en reprenant les différentes tâches et les différents intervenants impliqués dans la mise en place du logiciel Collec-Science au sein du laboratoire de Chrono-environnement de l'université de Bourgogne-Franche Comté.

Business Model Canvas – Appropriation du Logiciel Collec-Science

<table>
<tr>
<td rowspan="2">Partenaires clés
• (A) : E. Quinton – EABX IRSTEA
• (B) : C. PLumejeaud UMR LIENSS/ RZA</td>
<td>Activités clés
• Analyse des besoins (A,B,C,D,E)
• Appui en développement (E)
• Administration de la plate-forme de production (F)
• Identification des personnes concernées par le logiciel au sein du laboratoire (E, C, D)
• Animation auprès des techniciens en charge de la gestion des échantillons (E)
• Définition des modèles d'échantillons (E, G1 et G2)</td>
<td rowspan="2">Offre (proposition de valeur)
Gestion des échantillons du laboratoire
• D'un point de vue informatique :
o centralisation des informations
o accès facilité
o gain de temps
• D'un point de vue description des échantillons :
o organisation
o homogénéisation</td>
<td>Relation client
• Formation des utilisateurs (E)
• Assistance aux utilisateurs (E, G1, H et I)
• Proposition de demandes d'évolution du logiciel en fonction des besoins des utilisateurs (E)</td>
<td rowspan="2">Segments de clientèle
• Ingénieurs/ Techniciens gérant les échantillons (stockage physique, répartition des échantillons dans les lieux de stockages, ...) (G)
• (H) : Scientifiques ayant besoin de pouvoir rechercher facilement des échantillons
• (I) : Gestionnaire du laboratoire pour la gestion des lieux de stockage (intervention en cas de panne, ...)</td>
</tr>
<tr>
<td>Ressources clés
• (C) : F. Raoul – directeur adjoint CE
• (D) : S. Damy – responsable de l'axe transversal Données - Chef de projet
• (E) : A. Maindron – développeur et animateur du projet
• (F) : JD. Tissot et CH. Falconnet - Admin système
• (G) : Ingénieurs / Techniciens gérant les échantillons (G1 : référents, G2 : utilisateurs)</td>
<td>Canaux de distribution
• Communication interne (C,D,E)
• Réunions organisées par l'animateur du projet (E) avec les techniciens (G1) et les utilisateurs finaux en consultation (G2, H, I)</td>
</tr>
<tr>
<td colspan="3">Structure des coûts
• Coûts de personnels :
o 6 mois d'Ingénieur d'études (E)
o 10 % E/C sur 6 mois (D), 5 % E/C sur 6 mois (C)
o 5 % à 15 % ingénieurs/techniciens sur 6 mois</td>
<td colspan="3">Sources de revenus
• Gestion des échantillons de qualité - Appui à la recherche :
o optimisation du temps de gestion des échantillons
o mutualisation et fiabilisation des données concernant les échantillons
• Publications scientifiques (posters, articles, communications)</td>
</tr>
</table>

Figure 3. Business Model Canvas correspondant à la mise en place de Collec-Science dans le laboratoire de Chrono-Environnement.

Compte-tenu de la nature des activités du laboratoire et de l'organisation des espaces de stockage, l'informatisation de la gestion des échantillons imposait une vision globale, et donc la mobilisation de l'ensemble des acteurs dès les premières étapes. Le projet a pu être porté par un animateur recruté pendant six mois, qui a en outre

collaboré au développement du logiciel et l'a fait évoluer en fonction des besoins spécifiques locaux. Le recours à des référents externes (le développeur et l'animatrice des Zones-Ateliers pour la gestion des échantillons) a été nécessaire pour faciliter la compréhension des concepts manipulés par le logiciel et appréhender sereinement son paramétrage. Bien que cela n'apparaisse pas dans le *Business Model Plan*, la cheffe de projet souhaiterait pouvoir basculer l'hébergement de l'application vers une plateforme externalisée, les ressources informatiques au sein de son laboratoire rendant la gestion technique de celle-ci peu pérenne.

3.3. Diffusion du logiciel Collec-Science

Le troisième cas d'utilisation que nous avons étudié correspond à l'ensemble des tâches à réaliser pour diffuser le logiciel auprès d'autres laboratoires. Dès la publication des premières versions, nous avons ressenti le besoin de mettre en place des outils de communication et d'animation de la communauté des utilisateurs. En nous appuyant sur le livre « Logiciels et objets libres. Animer une communauté autour d'un projet libre » (Ribas *et al*., 2016), nous avons mis en place un site web, des plates-formes de démonstration, des listes de diffusion, défini des règles d'organisation du développement, etc.

Malgré un appui fort réalisé par le réseau des Zones Ateliers, différents freins ont été identifiés, notamment en ce qui concerne la mise en place de la plate-forme technique d'hébergement du logiciel. Celui-ci nécessite la mise en place d'un serveur web Apache [6], des couches logicielles associées (PHP, JAVA, etc.) et d'un serveur de bases de données Postgresql [7]. Les premières versions nécessitaient la réalisation d'une suite d'opérations manuelles pour installer tous les composants, et peu de laboratoires disposaient en interne des ressources nécessaires pour réaliser ce travail, soit par manque de temps, soit par manque de compétences. Des scripts permettant une installation quasi-automatique du logiciel ont été mis au point. Deux stratégies différentes ont été mises en œuvre. La première permet de déployer une image dans un container Docker [8], la seconde déploie automatiquement les composants nécessaires dans un serveur Linux Debian [9]. La mise en place de ces scripts n'a pas permis de répondre à toutes les attentes : certains laboratoires n'avaient pas la capacité technique ou humaine suffisante pour déployer la solution en interne. Dans le cadre de l'animation des Zones Ateliers, une machine virtuelle a été louée auprès d'INRAE pour héberger les instances des laboratoires, membres d'une Zone Atelier, qui le souhaitaient.

L'ensemble des acteurs mobilisés et des tâches réalisées pour assurer le développement, la diffusion et la maintenance du logiciel sont récapitulés dans la figure 4.

6. http://httpd.apache.org/
7. https://www.postgresql.org/
8. https://github.com/jancelin/docker-collec
9. https://github.com/Irstea/collec/raw/master/install/deploy_new_instance.sh

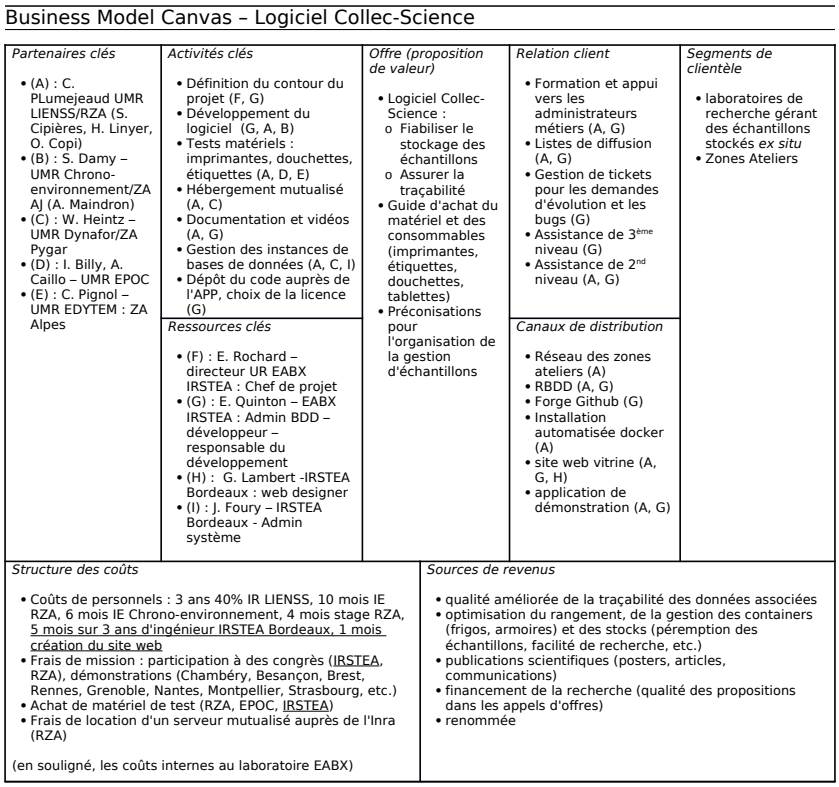


Figure 4. Business Model Canvas correspondant à la création et à la diffusion du logiciel Collec-Science

Compte-tenu de la complexité des interactions entre les différents intervenants et les actions menées, nous avons représenté dans la figure 5 les principales activités réalisées.

Ce diagramme met en exergue trois rôles essentiels : le développeur, le chef de projet pour la diffusion, et le responsable métier (ou le chef de projet) de l'organisme déployant le logiciel. Ces rôles sont génériques : ils peuvent être portés par une ou plusieurs personnes différentes. Le développeur s'occupe à la fois du codage, de la gestion des versions, du traitement des demandes d'améliorations ou des signalements de dysfonctionnements, de la rédaction de la documentation technique. Le chef de projet pour la diffusion prend en charge toute la partie communication, appui, formation, rédaction de la documentation à destination des utilisateurs, etc. Le responsable métier de l'organisme déployant le logiciel a un rôle de chef de projet pour sa mise en œuvre dans son laboratoire, en allouant ou en obtenant les moyens techniques, humains et

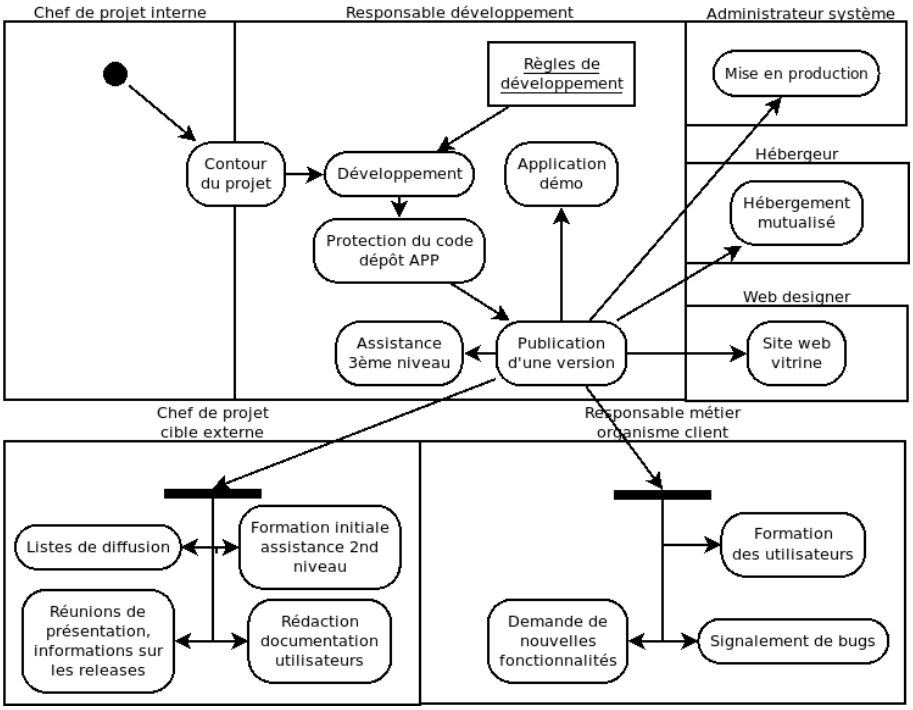


Figure 5. Diagramme d'activité représentant les acteurs impliqués et les tâches menées pour la création et la diffusion du logiciel Collec-Science

financiers nécessaires, en supervisant l'ensemble des opérations et en s'assurant de la qualité des données injectées dans le système.

4. Discussion

4.1. Identification des rôles à mobiliser

La notion de rôle est importante pour identifier clairement la position d'une personne dans une organisation. Plusieurs types de rôles peuvent être identifiés : axés sur les tâches, sur les relations ou sur l'autonomie (Zhu *et al*., 2006). Nous nous attachons ici à identifier des rôles axés sur des tâches relevant du même domaine ou de même responsabilité. Quel que soit le mode de mise en œuvre d'un logiciel, qu'il soit développé en interne ou récupéré depuis un autre organisme, trois rôles sont absolument indispensables : un chef de projet, qui allouera les moyens adéquats et délimitera le projet, un administrateur de données / curateur de données / animateur du projet, qui le fera vivre au quotidien, et un administrateur des systèmes d'informations, qui s'occupera de gérer la plate-forme technique. Si cette dernière fonction ne peut être mobilisée, le recours à une offre d'hébergement devient indispensable.

Dans l'ensemble des laboratoires que nous avons accompagnés, le rôle du **curateur de données** nous a semblé prépondérant. Celui-ci doit disposer de compétences variées dans les domaines de la gestion de l'organisation (structuration et mise en place de collaborations, de politiques internes, etc.), du service aux utilisateurs (accès et réutilisation des données, formation, etc.) et de la gestion des données proprement dites ou de la technologie (Tammaro *et al*., 2014). Sa connaissance profonde de tous les aspects liés aux données en font un interlocuteur essentiel pour mettre en place une gestion des données, et particulièrement dans le domaine des collections, celui que nous avons traité. Toutefois, cette fonction est peu présente dans les unités de recherche. En novembre 2019, sur 39 postes ouverts au concours du CNRS pour la filière *informatique, statistiques et calcul scientifique*, un seul concernait la fonction d'administrateur de bases de données, contre 15 en ingénierie logicielle et 22 en calcul scientifique. Pourtant, les besoins dans ce domaine augmentent régulièrement, notamment avec les problématiques liées à l'*Open Data*, la mise en place de plans de gestion de données qui sont de plus en plus demandés dans le cadre des projets de recherche H2020 européens [10], ou pour répondre aux exigences du protocole de Nagoya [11].

Pour la diffusion d'un logiciel, deux rôles deviennent prédominants : le développeur et le chef de projet de diffusion. Le **développeur** doit non seulement créer le logiciel, mais également assurer son suivi avec la gestion des tickets, la rédaction de la documentation, l'assistance de troisième niveau, la mise en place de scripts de mise à jour, etc. Le **chef de projet de diffusion** a un rôle majeur en matière de formation des chefs de projets locaux, d'assistance de second niveau, de communication, etc. Son rôle se rapproche de celui des revendeurs de solutions dans le contexte du PLM (*Product Lifecycle Management*). Restuccia *et al*. (2016) ont identifié quatre rôles que peuvent jouer ces intermédiaires : informateurs de problèmes, conseillers en solutions, metteurs en œuvre de solutions, et gestionnaires de solutions. Dans la plupart des cas, le chef de projet de diffusion a eu un rôle de conseil, mais a pu aider parfois à implémenter la solution. Un recrutement de quelques mois d'un ingénieur dédié à cet aspect a d'ailleurs été réalisé par le réseau des Zones Ateliers.

Dans le tableau 2, nous avons cherché à identifier les différents rôles à mobiliser, en fonction du cas d'utilisation : écriture en interne d'un logiciel, création et diffusion d'un logiciel vers des structures tierces, et appropriation d'un logiciel fourni par une autre structure.

Plusieurs rôles peuvent être assumés par une même personne, certains pourraient être répartis au sein d'une équipe ou d'un collectif. Dans l'appui que nous avons apporté lors du déploiement de Collec-Science, nous avons identifié que l'absence d'une

10. http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

11. Protocole de Nagoya sur *l'accès aux ressources génétiques et le partage juste et équitable des avantages découlant de leur utilisation relatif à la convention sur la diversité biologique*, publié en 2012 (https://www.cbd.int/abs/doc/protocol/nagoya-protocol-fr.pdf)

Tableau 2. Rôles identifiés lors de la mise en œuvre d'un logiciel de gestion de données

Rôle	Description	Interne	Diffusion	Appropriation
Directeur	Ancrage du projet dans la pratique du laboratoire	X	X	X
Chef de projet	Pilotage du projet	X	X	X
Administrateur des données	Responsable de la qualité des données, animateur du projet	X	X	X
Responsable du développement	Supervision, mise en production des nouvelles versions et tâches liées à la protection du logiciel		X	
Développeur	Ajout de fonctionnalités, corrections	X	X	(X)
Administrateur système	Maintien en condition opérationnelle des plates-formes techniques	X	X	X
Hébergeur	Gestion de la plate-forme d'hébergement des instances des laboratoires clients		X	
Chef de projet diffusion	Toutes tâches liées à la mise en œuvre du logiciel dans des structures tierces		X	
Web designer	Conception du site web vitrine		X	
Responsables applicatifs	Utilisateurs avancés du logiciel	X		X
Utilisateurs	Utilisateurs du logiciel	X		X

personne pour tenir un des rôles était un facteur important de retard dans la mise en œuvre du projet, voire pouvait conduire à son échec.

4.2. Identification des tâches principales pour déployer un logiciel

Une fois que le laboratoire a décidé de mettre en œuvre un logiciel dont il n'a pas géré le développement, le projet commence en général par la mise à disposition des moyens techniques, humains et financiers nécessaires, puis par une communication vers les équipes locales, sous des formes variées. Le déploiement du logiciel proprement dit intègre des tâches de paramétrage, de formation des utilisateurs, d'assistance. Outre celles-ci, la gestion de la plate-forme technique apparaît comme une réelle problématique, à la fois pour sa mise en place, son administration (réalisation des sauvegardes, etc.) ou pour appliquer les mises à jour. Dans la pratique, peu de laboratoires disposent des ressources techniques suffisantes pour gérer ce type de plates-formes, ainsi des **solutions de mutualisation** deviennent quasiment incontournables. Mais, comme pour les solutions de type *Cloud* dont les risques ont été largement étudiés

(Neumann, 2014), des précautions doivent être prises pour garantir la sécurité des données et la réversibilité de l'hébergement (garantie de sauvegarde, base de données dédiée susceptible d'être récupérée intégralement, *p. e.*).

4.3. Identification des tâches principales impliquées dans la diffusion du logiciel Collec-Science

Outre les classiques opérations de communication (site web, interventions dans différents séminaires, etc.), nous avons identifié plusieurs tâches essentielles à la diffusion du logiciel, dans deux domaines principaux : l'appui aux chefs de projets locaux et la mise en place de solutions facilitant la diffusion technique.

Les chefs de projets locaux ont, dans la plupart des cas, réclamé un appui personnalisé, que nous avons apporté soit par l'intermédiaire de formations, soit par des échanges directs. La mise au point de **supports de communication** différents, comme des vidéos explicitant certains points d'utilisation particuliers, a été également appréciée.

Concernant la diffusion technique, la fourniture de **scripts d'installation quasi-automatiques** a été demandée très tôt. La mise en place d'une **plate-forme d'hébergement mutualisée** a été un facteur de facilitation fort.

Enfin, l'animation de la communauté des utilisateurs prend de plus en plus d'importance au fil du temps. En 2019, soit trois ans après la sortie de la première version du logiciel Collec-Science, un comité de pilotage du logiciel a été mis en place. Il est constitué de représentants de divers laboratoires soit en phase d'appropriation, soit en cours d'utilisation. Ce comité, organisé en groupes de travail sur différentes thématiques, doit faciliter l'organisation de formations, les échanges sur les pratiques et besoins des utilisateurs. Regroupant des personnes de différents laboratoires (répartis sur toute la France), il s'intéresse aussi à l'obtention de financements pour l'animation, le suivi du logiciel et des actions plus ponctuelles ou plus ciblées telle que la location de la plate-forme mutualisée.

4.4. Pérennisation du projet

La diffusion d'un logiciel Open-Source s'appuie sur deux rôles essentiels : le développeur et le chef de projet diffusion. La question de la pérennisation de ces deux rôles se pose. S'il est relativement facile de les mettre en place au début du projet, ils doivent s'inscrire dans la stratégie des structures qui les portent. Les investissements consentis pour diffuser le logiciel doivent pouvoir induire des bénéfices, immatériels ou non. Nous en avons identifié deux types. Le premier concerne l'amélioration du logiciel suite aux demandes des divers utilisateurs externes, les corrections de bogues ou les nouvelles fonctionnalités implémentées profitant directement au laboratoire d'ori-

gine [12]. Le second est un impact en terme de notoriété, soit par la publication d'articles ou la participation à des congrès, soit par la reconnaissance du laboratoire ou des personnes qui portent le projet, ce qui peut avoir un effet indirect pour l'obtention de programmes de recherche. Mais cela reste insuffisant. Dès lors que le logiciel se diffuse, d'autres moyens doivent être trouvés. En effet, le laboratoire d'origine vise l'innovation par la recherche et consacre ses forces à sa mission première et non pas à assumer la maintenance et la diffusion d'un logiciel. Lisein *et al*. (2009) ont étudié plusieurs sociétés éditant des logiciels Open-Source, et la question de leur pérennisation est au cœur de leurs préoccupations. Une des sociétés étudiées fonctionnait uniquement avec des subventions et du mécénat, et sa survie n'était pas assurée sur le long terme. La longévité d'un projet applicatif dans le domaine de la recherche reste donc une question encore délicate.

5. Conclusions et perspectives

L'utilisation d'un logiciel de gestion de données de recherche, conçu en interne ou par un autre laboratoire est un processus qui s'inscrit dans le mouvement pour la science ouverte. Elle permet de répondre aux besoins d'ouverture et de partage des données. Sa diffusion et sa réutilisation permettent des économies de moyens et d'échelle, mais cela impose la réalisation de tâches spécifiques et la mobilisation de profils particuliers, tant pour les concepteurs de l'application que pour ceux qui le mettent en œuvre. L'utilisation de la grille d'analyse *Business Model Canvas* nous a permis de définir clairement les coûts et les gains associés à ces processus, et d'identifier les rôles mobilisés. Cet exercice assez peu habituel dans la milieu de la recherche nous a permis de mettre en avant, vis à vis de nos structures d'accueil, l'apport et l'intérêt de ces logiciels qui, dans le contexte actuel d'ouverture et partage des données de la recherche, sont devenus indispensables, , mais aussi leurs coûts en termes de ressources humaines notamment.

Dans ce travail, nous avons identifié des rôles clés qui facilitent la mise en œuvre d'une informatisation de gestion de données : le chef de projet, le curateur de données / animateur du projet, l'administrateur des systèmes d'information. Quand le logiciel est diffusé, un nouveau rôle prépondérant émerge : le chef de projet pour la diffusion.

La diffusion d'un logiciel comme Collec-Science en *Open-Source* ne peut se suffire d'une simple publication du code dans une plate-forme ouverte. Elle impose la mise en place d'une plate-forme d'hébergement mutualisée, de la communication sur des supports divers, et des actions d'appui vers les animateurs locaux, soit sous forme de formations spécifiques, soit sous forme d'appui personnalisé. Si, jusqu'à présent, nous avons pu répondre aux besoins des laboratoires tiers dans leur mise en place du logiciel Collec-Science, nous nous interrogeons quant à la pérennisation de cet appui,

12. La dernière version de Collec-Science intègre une gestion de méta-données et une représentation cartographique de l'emplacement des échantillons d'une collection, deux demandes issues des laboratoires partenaires, et qui ont trouvé une utilisation immédiate dans les projets du laboratoire d'origine.

au moins à moyen terme, avec deux questions principales : comment garantir un hébergement de qualité, et comment apporter l'aide de second niveau ? Pour les logiciels commerciaux, ces tâches sont souvent prises en charge par les revendeurs, qui peuvent proposer des niveaux d'intervention plus ou moins importants (Restuccia *et al*., 2016). Dans le contexte de l'Open-Source, la gestion d'une communauté peut être une des réponses, mais celle-ci ne peut fonctionner qu'avec la présence d'acteurs et d'animateurs (Dupont *et al*., 2017), qu'il faut coordonner. L'attribution des responsabilités et le parrainage (avec les financements qui vont avec) sont des facteurs essentiels pour assurer la réussite d'un projet (Viseur, 2013).

Il n'est pas interdit d'envisager un recours à des structures privées pour gérer ces aspects. Si le choix d'une licence *Open-Source* pour diffuser le logiciel est impératif dans le contexte de la recherche publique (Quinton *et al*., 2018), diverses stratégies ont été proposées pour monnayer et rendre financièrement intéressant les logiciels (Shahrivar *et al*., 2018 ; Osterwalder, 2004 ; Osterwalder, Pigneur, 2011). Certaines pistes pourraient s'appliquer à nos cas de figure : facturation de l'hébergement pour les structures qui ne peuvent déployer la solution en interne, prestation d'appui au démarrage comprenant l'analyse des besoins propres et la configuration initiale, abonnement pour un accès à une maintenance de second niveau pour les responsables métiers, voire de premier niveau pour les utilisateurs, etc. Toutefois, la volumétrie des déploiements reste trop faible pour qu'une structure puisse envisager de ne travailler qu'avec un seul logiciel : c'est tout un écosystème d'appui qu'il faudrait mettre en place autour de différents produits.

Une autre solution serait de s'appuyer sur une unité de service d'un établissement de recherche assurant les missions d'hébergement et d'aide à la configuration. Pour Collec-Science, la plate-forme mutualisée mise en place à INRAE, et financée par un budget alloué par le réseau des Zones-Ateliers, rentre partiellement dans ce schéma.

Bibliographie

Dupont L., Gabriel A., Camargo M., Guidat C. (2017, juin). Collaborative Innovation Projects Engaging open communities: a Case Study on Emerging Challenges. In *23rd ICE/IEEE International Conference on Technology, Management and Innovation.* Madeira Island, Portugal. Consulté sur https://dx.doi.org/10.1109/ICE.2017.8280002

Eberendu A. C. (2015). *Evaluation of Software Project Failure and Abandonment in Tertiary Institutions in Nigeria.* IISTE. Consulté sur https://citeseerx.ist.psu.edu/viewdoc/download ?doi=10.1.1.909.3042&rep=rep1&type=pdf

Fjeldsted A., Adalsteinsdottir G., Howard T., McAloone T. (2012, janvier). Open Source Development of Tangible Products - from a business perspective. Consulté sur https://www.researchgate.net/publication/261365095_Open_Source_Development_of _Tangible_Products_-_from_a_business_perspective

Lisein O., Pichault F., Desmecht J. (2009). Les business models des sociétés de services actives dans le secteur Open Source. *Systemes d'information management*, vol. Volume 14, n° 2, p. 7–38. Consulté sur https://www.cairn.info/revue-systemes-d-information-et -management-2009-2-page-7.htm (Publisher: ESKA)

Luoma E., Rönkkö M., Tyrväinen P. (2012). Current Software-as-a-Service Business Models: Evidence from Finland. In M. A. Cusumano, B. Iyer, N. Venkatraman (Eds.), *Software Business*, p. 181–194. Berlin, Heidelberg, Springer Berlin Heidelberg.

Mendoza A., Carroll J., Stern L. (2010, mars). Software appropriation over time: from adoption to stabilization and beyond. *Australasian Journal of Information Systems*, vol. 16, n° 2. Consulté sur https://journal.acs.org.au/index.php/ajis/article/view/507

Neumann P. G. (2014, septembre). Risks and myths of cloud computing and cloud storage. *Commun. ACM*, vol. 57, n° 10, p. 25–27. Consulté sur http://doi.acm.org/10.1145/2661049

Osterwalder A. (2004). The Business Model Ontology – A Proposition in a Design Science Approach. *Doctorat en Informatique de Gestion. Ecole des Hautes Etudes Commerciales de l'Université de Lausanne.*. Consulté sur https://www.researchgate.net/publication/33681401_The_Business_Model_Ontology_-_A_Proposition_in_a_Design_Science_Approach

Osterwalder A., Pigneur Y. (2011). *Business Model nouvelle génération*. Pearson France. Consulté sur https://www.pearson.fr/FR/book/?GCOI=27440100730760

Plumejeaud-Perreau, Quinton E., Pignol C., Linyer H., Ancelin J., Cipière S. *et al*. (2019, août). Towards better traceability of field sampling data. *Computers & Geosciences*, vol. 129, p. 82–91. Consulté sur https://doi.org/10.1016/j.cageo.2019.04.009

Quinton E., Plumejeaud-Perreau C., Linyer H., Ancelin J., Pignol C., Cipière S. *et al*. (2018, mai). Sample management for scientific research with Collec-Science. In *INFORSID*, p. 41-61. Nantes, France. Consulté sur https://hal.archives-ouvertes.fr/hal-01825250 (INFORSID, Nantes, FRA, 28-/05/2018 - 31/05/2018)

Restuccia M., Brentani U. de, Legoux R., Ouellet J.-F. (2016). Product life-cycle management and distributor contribution to new product development. *Journal of Product Innovation Management*, vol. 33, n° 1, p. 69-89. Consulté sur https://onlinelibrary.wiley.com/doi/abs/10.1111/jpim.12261

Ribas S., Guillaud P., Ubeda S. (2016). *Logiciels et objets libres. animer une communauté autour d?un projet libre* (Framasoft, Ed.). Consulté sur https://framabook.org/logiciels-et-objets-libres/

Shahrivar S., Elahi S., Hassanzadeh A., Montazer G. (2018, novembre). A business model for commercial open source software: A systematic literature review. *Information and Software Technology*, vol. 103, p. 202–214. Consulté sur https://doi.org/10.1016/j.infsof.2018.06.018

Tammaro A. M., Ross S., Casarosa V. (2014). Research Data Curator: the competencies gap. In *BOBCATSSS 2014 Proceedings.* Consulté sur https://proceedings.bobcatsss2014.hb.se/article/view/327

Viseur R. (2013, juin). Identifying Success Factors for the Mozilla Project. In E. Petrinja, G. Succi, N. Ioini, A. Sillitti (Eds.), *9th Open Source Software (OSS)*, vol. AICT-404, p. 45-60. Koper-Capodistria, Slovenia, Springer. Consulté sur https://dx.doi.org/10.1007/978-3-642-38928-3_4 (Part 1: Full Papers - Innovation and Sustainability)

Zhu H., Zhou M., Seguin P. (2006, Nov). Supporting software development with roles. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 36, n° 6, p. 1110-1123. Consulté sur https://doi.org/10.1109/TSMCA.2006.883170

Ingénierie des processus

Analyse conceptuelle des processus métier sensibles - *Mariam Ben Hassen, Mohamed Turki and* Faïez *Gargouri* (article long)

Un cadre méthodologique As-Is/As-If pour guider le développement des méthodes d'évolution continue - *Ornela Cela, Mario Cortés-Cornax, Agnès Front et Dominique Rieu* (résumé étendu)

Tiers-Lieu pour les services d'information : la valeur de la modélisation conceptuelle - *Jolita Ralyté and Michel Léonard* (résumé étendu)

Analyse conceptuelle des processus métier sensibles

Mariam Ben Hassen, Mohamed Turki, Faïez Gargouri

Université de Sfax, Institut Supérieur d'Informatique et de Multimédia de Sfax, Laboratoire de recherche MIRACL, P.O. Box 242, 3021 Sfax, Tunisie

mariem.benhassen@isims.usf.tn, mohamed.turki@isims.usf.tn, faiez.gargouri@isims.usf.tn

RÉSUMÉ. *Afin d'améliorer leurs performances, les organisations modernes sont de plus en plus conscientes de la nécessité d'identifier, préserver et gérer les connaissances cruciales (individuelles et collectives) mobilisées par leurs processus métier sensibles (Sensitive Business Processes : SBP). Il va falloir, ainsi, caractériser, identifier, spécifier, modéliser et analyser ces processus, afin d'optimiser les activités d'identification, de localisation et de gestion des connaissances sur lesquelles il faut capitaliser. Ce papier introduit la problématique de modélisation des SBP. Notre objectif consiste à mener une analyse conceptuelle se rapportant à la notion de SBP. Nous proposons, en premier lieu, une caractérisation rigoureuse des SBP (qui le distinguent des BP classiques, structurés et conventionnels). En second lieu, nous proposons une classification multidimensionnelle des aspects et des exigences spécifiques de modélisation des SBP pour obtenir des modèles expressifs, complet et rigoureux. Sur la base de ces exigences, nous présentons, en dernier lieu une analyse comparative des différentes approches et langages de modélisation actuels pour en déduire leur expressivité et leur capacité à représenter explicitement les caractéristiques pertinentes de SBP.*

ABSTRACT. *In order to improve their performance, modern organizations are increasingly aware of the need to identify, preserve and manage the crucial knowledge (individual and collective) that are mobilized by their sensitive business processes (SBPs). Thus, it will be necessary to characterize, identify, specify, model and analyze these processes, in order to optimize the activities of identification, localization and management of the knowledge on which it is necessary to capitalize. This paper introduces the problematic of the SBP modeling. Our objective is to provide a conceptual analysis related to the concept of SBP. First of all, we propose a rigorous characterization of SBP (which distinguishes it from classic, structured and conventional BPs). Secondly, we propose a multidimensional classification of SBP modeling aspects and requirements to develop expressive, comprehensive and rigorous models. Finally, we present an in-depth study of the different modeling approaches and languages, in order to analyze their expressiveness and their ability to perfectly and explicitly represent the new specific requirements of SBP modeling.*

MOTS-CLÉS : Gestion des connaissances, processus métier sensible, modélisation des processus métier, langages et approches de modélisation des processus

KEYWORDS : Knowledge management, Business process modeling, Sensitive business process, Modeling approaches and languages

1. Introduction

De nos jours, la modélisation des processus métier sensibles (SBP) est devenue une préoccupation primordiale de toute organisation performante pour la gestion efficace de leur patrimoine de connaissances. En effet, plus les BP sont sensibles, plus ils sont susceptibles de mobiliser des connaissances cruciales. Ainsi, une meilleure identification et modélisation de ce type de processus peuvent optimiser l'identification et la gestion des connaissances sur lesquelles il faut capitaliser.

L'intégration de la gestion des connaissances (KM) avec l'ingénierie des processus métier est rapidement devenue la tâche pratique et théorique la plus pressante et la plus prometteuse dans le domaine de KM. Dans ce contexte, plusieurs tentatives ont déjà été élaborées pour intégrer/coupler le domaine de KM avec le domaine de modélisation des processus métier (BPM). Nous distinguons principalement deux classes d'approches : (i) des approches de KM orientées processus (KM-BPM) (*e.g.,* (Abecker, 2001), (Gronau et al., 2005), (Woitsch and Karagiannis, 2005)), et (ii) des approches de BPM orientées connaissances (BPM-KM) (*e.g.,* (Papavassiliou and Mentzas, 2003), (Heisig, 2006), (Strohmaier et al., 2007), (Supulniece et al., 2010), (Bušinska and Kirikova, 2011), (Sultanow et al., 2012), (Netto et al., 2013), (Ouali et al., 2016)). Pour la première classe de méthodes, il s'agit de considérer les BP comme un sujet de KM et d'intégrer le concept « processus » dans le processus de KM ainsi que dans les approches et notations de modélisation des connaissances. Tandis que la deuxième classe de méthodes considère les BP comme un point initial pour KM, qui intègre le concept ou la dimension « connaissance » dans les langages de BPM. Cependant, cette intégration cruciale de domaines de BPM-KM n'a pas encore reçu suffisamment d'attention et n'est pas étudiée en profondeur. Notamment, la dimension « connaissance » (*e.g.,* les connaissances collectives, les connaissances explicites, les connaissances tacites et les sources de connaissances qui sont mobilisées par les activités organisationnelles, les différentes possibilités de conversion de connaissances, etc.), n'est pas explicitement et complétement représentée, intégrée et implémentée dans les modèles de BP/SBP comme étant une dimension centrale de modélisation des BP/SBP (Ben Hassen et al., 2017b ; 2017 c ; 2018).

Par ailleurs, dans la revue de la littérature, il existe peu de travaux dans le domaine de KM-BPM, s'intéressant à la délimitation du champ des connaissances sur lesquelles il faut capitaliser. Notamment, les différentes méthodes de repérage des connaissances centrées sur l'analyse des processus visent à identifier, modéliser et analyser les processus sensibles, afin d'identifier les connaissances cruciales ((Grundstein, 2009), (Saad et al., 2009), (Turki et al., 2014 ; 2016), (Ghrab and Saad, 2016), (Ghrab et al., 2017)). Cependant, l'activité critique de « modélisation des processus métier sensibles » n'est pas ni explicitée ni étudiée en profondeur. Or, cette opération constitue une phase primordiale du processus d'identification des connaissances cruciales dans les organisations. En effet, certaines limitations peuvent être recensées, dont nous citons particulièrement: (i) absence d'une description et une caractérisation fine de la notion de « processus métier sensible (SBP) » qui intègre, notamment, la « dimension connaissance », (ii) absence d'une spécification conceptuelle rigoureuse de ce type de processus, (iii) la non prise en compte des approches et des langages de BPM appropriés pour la modélisation des SBP qui supportent parfaitement et complétement ses caractéristiques spécifiques, et (iv) absence d'une approche scientifique rigoureuse pour la spécification et la modélisation des SBP dans une perspective de gestion des connaissances cruciales. Par ailleurs, nous notons l'absence de définitions rigoureuses et consensuelles qui caractérisent les SBP. Ainsi, il n'existe pas d'approches scientifiques et des modèles formels permettant de spécifier et de modéliser ce

type de processus. Ces différentes lacunes mènent, évidement, à développer des modèles de SBP incomplets, ambigus et incompréhensibles (Ben Hassen et al., 2017a ; 2017b ; 2018).

Afin de remédier à ces insuffisances recensées, enrichir et améliorer la modélisation des SBP, nous proposons au premier abord dans ce papier, une analyse conceptuelle des SBP. Concrètement, l'objectif de ce papier est de : (i) proposer une caractérisation rigoureuse des SBP ; (ii) proposer une classification multidimensionnelle des aspects et des exigences de modélisation des SBP ; (iii) mener une analyse des différentes approches et langages de modélisation actuels proposés dans le domaine de BPM-KM, pour en déduire leur expressivité et leur capacité à représenter explicitement les caractéristiques et les nouvelles exigences de modélisation des SBP. Le reste de ce papier est organisé comme suit : la section 2 se focalise sur la description des SBP. D'abord, nous abordons la notion de SBP ainsi que ses principales caractéristiques. Puis, nous présentons les différentes dimensions de SBP reflétant ses aspects pertinents. Ensuite, nous présentons les dimensions et les exigences spécifiques de modélisation des SBP. Par la suite, nous étudions, les différentes approches et langages susceptibles de modéliser et de représenter ce type de processus. Nous clôturons ce papier par une conclusion en donnant les perspectives de ce travail de recherche.

2. Notion de processus métier sensible (Sensitive Business Process : SBP)

Dans notre étude, nous nous intéressons à un type spécifique de processus, à savoir, le processus métier sensible (SBP). L'identification et la modélisation de ces processus représentent des opérations primordiales dans le processus de gestion des connaissances cruciales dans les organisations.

2.1. Définition des processus métier sensibles

Il existe nécessairement au sein de toute organisation, des processus plus sensibles que d'autres. Cette sensibilité peut se mesurer par les enjeux considérables que représente un processus donné pour l'organisation.

Peu de définitions ont été proposées dans la littérature pour le concept de processus métier sensible ((Grundstein, 2009), (Turki et al., 2016), (Ben Hassen et al., 2017a ; 2018)).

Selon (Grundstein, 2009), « *un processus sensible est un processus présentant des enjeux reconnus collectivement : faiblesse du processus présentant le risque de ne pas atteindre les objectifs de coût, de délai, de qualité requis pour la production des biens ou des services ; obstacles importants à surmonter ; challenges difficiles à atteindre ; biens ou services produits qui sont stratégiques pour l'entreprise* ».

D'après (Turki et al., 2016), un processus d'organisation sensible est défini comme étant « *un processus susceptible de mobiliser des connaissances sur lesquelles il faut capitaliser* ». Ainsi, le choix du SBP s'opère par rapport à l'impact de certains critères jugés stratégiques pour l'organisation et pour le processus. Ces critères sont relatifs, par exemple, au degré de contribution du processus à l'atteinte des objectifs stratégique de l'organisation, à la complexité de la structure du processus, le nombre des activités critiques, l'affiliation des acteurs impliqués, la durée et le coût des processus.

Dans le cadre de nos travaux de recherche, nous nous appuyons sur la définition proposée par Turki et al. Ainsi, nous fournissons une nouvelle définition précise et plus complète qui capture les principaux éléments distinctifs d'un SBP et qui intègre tous les enjeux pertinents liés à l'intersection du domaine de BPM et de KM (notamment la

dimension connaissance). Quant à nous, « *Un processus métier sensible est un type particulier de processus métier centré sur les connaissances, les informations et les données. Il comprend un nombre élevé d'activités critiques (individuelles et/ou collectives), mobilisant des connaissances cruciales, sur lesquelles il faut capitaliser. Également, il contient des activités à forte intensité de connaissances qui valorisent l'acquisition, le stockage, la dissémination, le partage, la conversion et la création des connaissances individuelles et collectives (tacites et explicites). Ainsi, il mobilise une grande diversité de sources de connaissances consignant une masse très importante de connaissances hétérogènes. Son exécution implique la collaboration et l'interaction de nombreux participants (qui peuvent êtres internes et/ou externes à l'organisation) et dépend fortement des connaissances tacites et stratégiques des experts ayant des niveaux d'expertise et de compétences hétérogènes. Ce type de processus peut être semi-structuré, structuré ou non structuré, possédant un degré élevé de complexité, de flexibilité et de dynamisme. Par ailleurs, le degré de contribution de ce processus pour atteindre les objectifs stratégiques de l'organisation et le coût de sa réalisation sont très importants* » (Ben Hassen et al., 2018).

Dans ce qui suit, nous présentons les principales caractéristiques de SBP, qui les différencient de BP structurés et conventionnels.

2.2. Caractérisation des processus métier sensibles

En effet, un SBP est un type particulier de processus métier (BP). Ainsi, il hérite et partage toutes les caractéristiques pertinentes de BP (Cf. (Turki et al., 2016)). En revanche, il possède ses propres caractéristiques qui le distinguent des BP classiques (Cf. Tableau 1). Compte tenu, d'une part, des éléments récurrents dans les définitions proposées dans la revue de la littérature ((Grundstein, 2009), (Saad et al., 2009), (Turki et al., 2014 ; 2016), et d'autre part, l'émergence de nouvelles formes des organisations modernes, nous avons dérivé sept caractéristiques clés représentatives de la notion de SBP (Ben Hassen et al., 2017a ; 2017b ; 2017d ; 2018). Un BP est qualifié de « sensible », si l'une des conditions suivantes est vérifiée :

- **C1−Un SBP est un BP à forte intensité de connaissances (centré sur les connaissances)**. Il mobilise des connaissances cruciales (Grundstein, 2009), *i.e.,* des connaissances très spécifiques sur lesquelles il est nécessaire de capitaliser en priorité. En d'autres termes, le risque de leur perte et le coût de leur (re) création sont considérés comme étant importants. Ainsi, leur degré de contribution à l'atteinte des objectifs organisationnels est très important. De même, leur durée d'utilisation est longue (Ben Hassen et al., 2017b ; 2018). Outre les connaissances cruciales, un SBP mobilise et produit différents types de connaissances (*e.g.,* des connaissances tacites, des connaissances explicites, des connaissances individuelles, des connaissances collectives, des connaissances organisationnelles, des connaissances procédurales, des connaissances stratégiques, des connaissances externes, etc.) (Ben Hassen et al., 2018). Par ailleurs, un SBP comprend des activités qui valorisent l'acquisition, le stockage, la dissémination, le partage, la création et la (ré) utilisation des connaissances. Ainsi, il mobilise une grande diversité de sources de connaissances (*e.g.,* des supports physiques de connaissances) capitalisant une masse très importante de connaissances (Ben Hassen et al., 2017a ; 2017b). De plus, il dépend strictement des connaissances tacites incarnées dans l'esprit des parties prenantes (des experts, des spécialistes, etc.). En réalité, ces connaissances sont en grande partie implicites, rarement explicitées et disséminées par les agents/experts qui les détiennent. Elles

sont ainsi difficiles à repérer, exploiter et valoriser par d'autres collaborateurs. En outre, ce type de BP se focalise sur la conversion dynamique des connaissances (*e.g.,* la socialisation, l'externalisation, etc.) (Ben Hassen et al., 2017a ; 2017b).

- **C2− Un SBP est dirigé par les activités critiques**. Il comporte un nombre élevé d'activités critiques ((Grundstein, 2009), (Saad et al., 2009), (Turki et al., 2014 ; 2016)) qui mobilisent des connaissances cruciales. Par ailleurs, une activité est qualifiée de critique si elle mobilise : (i) des connaissances et des informations sources de connaissances imparfaites (*i.e.,* manquantes, mal maîtrisées, incomplètes, incertaines, etc.) qui sont nécessaires à la résolution des problèmes déterminants (Grundstein, 2009); (ii) des connaissances importantes, hétérogènes, consignées sur diverses sources de connaissances qui peuvent être dispersées et manquent parfois d'accessibilité ; (iii) des expertises et/ou des connaissances rares détenues par un nombre très restreint d'experts ; (iv) des connaissances flexibles (détenues par des experts) ; (v) des connaissances organisationnelles tacites très importantes (liées aux compétences et expériences pratiques des experts) (Ben Hassen et al., 2017a; 2018).
- **C3− Un SBP est dépendant des données et des informations (sources de connaissances)**. Un SBP dépend des données et des informations (et leurs sources) dans la modélisation des flux entre les activités. L'échange de données et d'informations constituent la base de création, de transfert et d'utilisation des connaissances (Ben Hassen et al., 2017b).
- **C4− Un SBP est orienté collaboration et interaction (humaine)**. Il présente un degré élevé d'interactions collaboratives entre les participants. Il comprend un nombre élevé d'activités collaboratives (intra/inter-organisationnelle), qui mobilisent, échangent, partagent et génèrent de nouvelles informations et connaissances individuelles et collectives crées durant un ensemble d'interactions entre les agents communicants. Ces activités nécessitent des décisions complexes et rapides parmi de multiples stratégies possibles pour atteindre les objectifs organisationnels (Ben Hassen et al., 2017a). De plus, il mobilise un nombre élevé de domaine/compétences métier (en termes d'unités d'organisation internes et externes impliquées dans le processus). Son exécution implique de nombreux participants et l'assistance de nombreux experts ayant des niveaux d'expertise et de compétences élevés et hétérogènes (Turki et al., 2014).
- **C6− Un SBP est régi par des contraintes et des règles**. En effet, les participants de SBP peuvent être influencés par, ou peuvent devoir se conformer, aux contraintes et aux règles qui régissent la performance des actions organisationnelles et la prise de décision (Reichert and Weber, 2012).
- **C7− Un SBP est axé sur les objectifs**. Il est guidé par les intentions de l'agent pour atteindre ses objectifs (individuels ou collectifs). Ainsi, il possède un degré élevé de dynamisme dans le changement des objectifs qui lui est associés (dans un contexte de prise de décision). Le changement d'un objectif organisationnel aboutit à une nouvelle intention distale organisationnelle (Turki et al., 2016) et influence sur la prise de décision des experts (Ben Hassen et al., 2018). Par ailleurs, leur contribution à l'atteinte des objectifs stratégiques de l'organisation est très importante. Ainsi, les SBP représentent les processus essentiels qui constituent le cœur des activités de l'organisation (Turki et al., 2016), (Ben Hassen et al., 2018).

Par ailleurs, la durée et le coût de la réalisation de SBP (leur conception et leur exploitation) sont importants (plus le BP a une durée importante et un coût élevé, plus il est sensible).

Le Tableau 1 présenté ci-après illustre une comparaison détaillée entre les SBP et les BP en nous basant sur un ensemble de critères et dimensions de comparaison qui nous semblent utiles et pertinentes pour notre contexte d'étude, du fait que nous nous intéressons à la modélisation des SBP dans une perspective d'identification des connaissances.

Tableau 1. Comparaison entre les caractéristiques de SBP et BP (classique)

Critères	**SBP**	**BP (classique)**
Sensibilité au risque	Élevé	Faible
Complexité structurelle	Très complexe /Flexible ou très flexible adaptatif à des contextes spécifiques et de nombreuses exceptions	Simple/(moyennement) complexe Peu flexible/Fortement prévisible
Dimensions de modélisation	Six dimensions de modélisation : fonctionnelle, organisationnelle, connaissance, informationnelle, intentionnelle et comportementale	Quatre dimensions : fonctionnelle, comportementale, organisationnelle et informationnelle
Types d'activités	Actions individuelles et collectives/ activités organisationnelles critiques/ activités à forte intensité de connaissances/ activités organisationnelles collaboratives (intra/inter-organisationnelles)	Actions individuelles et collectives Activités privées/interne à un processus intra-organisationnel
Représentation des connaissances	Processus à forte intensité de connaissances cruciales/Diverses sources de connaissances/Conversion dynamique des connaissances	N'inclut pas le concept/la dimension connaissance
Interaction communicative	Des interactions communicatives humaines élevées/échange et partages intenses des connaissances entre les agents communicants	Faible/Échange d'informations (des messages) entre les agents (un nombre restreint)
Structure d'enchainement d'activités	Structuré, structuré avec exceptions ad hoc, faiblement/semi structuré ou non structuré	Structuré, structuré avec peu d'exceptions
Données et Informations	Distinction entre les flux de données et les flux d'information Diversité de sources d'informations	Pas de distinction entre les données et les informations (et leurs flux)
Flux de contrôle avancé	Patrons de coordinations avancés/Divers flux de connaissances	Patrons de coordination de base
Participants	Typologie diversifiée internes et/ou externes à l'organisation /Son exécution dépend fortement des experts ayant un niveau élevé d'expertise, de performance, de créativité et d'innovation	La créativité et l'innovation des experts sont non requises
Événements	Des événements externes et imprévisibles	Événements standards
Ressources (complexes)	Large mobilisation des ressources immatérielles/Connaissances souvent difficilement accessibles	Mobilisation des ressources matérielles
Contraintes et règles	Nombre élevé	Nombre faible
Objectif et intention	Objectifs prédéfinis ou définis progressivement	Objectif connu a priori et prédéfini
Type de modélisation	Modélisation déclarative	Modélisation impérative
Durée de réalisation	Importante/élevée	Moyenne/Courte
Coût	Élevé	Moyen/Faible

À l'instar des caractéristiques de spécification des SBP susmentionnées, nous soulignons que la représentation et l'organisation des connaissances impliquées dans les SBP ou l'élaboration des modèles complets de SBP est beaucoup plus complexe et difficile. Cela, nécessite alors, le choix et l'adoption d'une approche ou d'un langage de BPM approprié pour représenter parfaitement toutes les dimensions et les exigences de modélisation des SBP, afin d'améliorer l'identification et la gestion des connaissances mobilisées et produites par ces processus.

3. Dimensions et exigences de modélisation des processus métier sensibles

Dans cette section, nous proposons une classification multidimensionnelle des aspects et des exigences à prendre en compte lors de la modélisation d'un SBP, afin d'obtenir un modèle complet et rigoureux.

3.1. Dimensions de modélisation des processus métier sensibles

L'expressivité de la modélisation d'un SBP repose essentiellement sur six dimensions en interaction, décrivant chacune un point de vue différent de ce processus. Il s'agit de la *dimension fonctionnelle*, la *dimension comportementale*, la *dimension organisationnelle*, la *dimension informationnelle*, la *dimension intentionnelle* et la *dimension connaissance* (Ben Hassen et al., 2017a ; 2017b ; 2017c ; 2018). Nous avons étendu les travaux de l'état de l'art et nous avons amélioré, d'une part, les cinq premières dimensions de définition des BP qui sont typiquement orientées modélisation métier (Van der Aalst et al., 2003), (Saidani and Nurcan, 2009), (Heidari et al., 2013), (Jankovic et al., 2015), (Turki et al., 2016), (Ben Said et al., 2018)), en proposant de nouvelles caractéristiques et de nouveaux concepts invariants de SBP. D'autre part, nous avons introduit la nouvelle « dimension connaissance » afin de prendre en compte tous les aspects pertinents liés au domaine de KM. Soulignons que, la connaissance est reliée à l'action, elle est mise en œuvre dans l'action et elle est essentielle pour son fonctionnement (Grundstein, 2009). De plus, les connaissances sont mobilisées pour réaliser un BP, elles sont créées comme un résultat de leur exécution et elles sont exploitées, échangées et partagées entre les agents opérants dans le processus.

3.1.1. La dimension fonctionnelle

La dimension fonctionnelle est considérée comme étant la dimension pivot dans la définition d'un SBP, qui fait référence aux autres dimensions comportementale, organisationnelle, informationnelle, intentionnelle et connaissance. Généralement, la dimension fonctionnelle représente les éléments de base du processus qui sont réalisées durant l'exécution de BP (*e.g.,* action, activité, sous-processus, tâche et fonction). Pour notre contexte de modélisation des SBP, nous considérons en plus des aspects structurels, les aspects d'interactions communicatives (intra/inter-organisationnelles), ainsi que les aspects dynamiques relatifs à la conversion et la création des informations et des connaissances (*e.g.,* la perspective collective des actions, les activités critiques mobilisant des connaissances cruciales, les activités à forte intensité de connaissances, les actions de conversion des connaissances, etc.). Ces aspects sont utiles et nécessaires pour caractériser un SBP, due, par exemple, à sa nature dynamique ainsi qu'à l'échange et le partage intenses des connaissances entre les agents communicants (Ben Hassen et al., 2017a ; 2018).

3.1.2. La dimension organisationnelle

Généralement, la dimension organisationnelle met en relief les différents participants/agents invoqués dans l'exécution des éléments de SB selon les rôles qui leurs sont confiés (*e.g.,* un humain, une unité organisationnelle). En plus de cette typologie, nous considérons d'autres types d'entités agentives capables de réaliser les différents types d'actions de SBP, tels qu'un collectif, un expert, un groupe informel, une organisation, un acteur interne et un acteur externe. Ces entités agentives interagissent et communiquent durant la réalisation des SBP, en échangeant et partageant des informations et des connaissances spécifiques et en générant de nouvelles (Ben Hassen et al., 2017a ; 2017b).

3.1.3. La dimension informationnelle

Généralement, la dimension informationnelle décrit les objets de données et d'information, les différentes entités d'information (*e.g.,* les messages, les artefacts, les entrées, les sorties, les ressources, les évènements, etc.). Pour notre contexte de modélisation des SBP, nous considérons, en plus de ces entités, la distinction explicite entre les données et les informations et de les différentier également des connaissances dans la représentation des flux entre les activités. Nous considérons ainsi, les flux de données, les flux d'informations, les différents types d'information, les différentes sources d'information qui sont mobilisées, générées et/ou manipulées par les activités de SBP, les communications et les discours, ainsi que la dynamique de transfert et de création des informations au sein des activités (Ben Hassen et al., 2017b ; 2018).

3.1.4. La dimension comportementale

La dimension comportementale décrit l'ordonnancement et les paramètres d'exécution des activités d'un BP/SBP. Elle décrit les flux de contrôle, les conditions de déclenchement et d'achèvement des éléments à exécuter, ainsi que leur coordination (à l'aide de patrons de coordination (Van der Aalst et al., 2003). En plus de ces éléments, nous considérons les flux de données, les flux d'informations ainsi que les flux de connaissances entre les différentes sources de connaissances et entre les activités de SBP (Ben Hassen et al., 2018).

3.1.5. La dimension intentionnelle

La dimension intentionnelle (appelée aussi la perspective contextuelle (Ben Said et al., 2018)) a été initialement proposée par (Saidani and Nurcan, 2009) pour décrir « le pourquoi » du processus en mettant l'accent sur les objectifs attendus du processus et les intentions à réaliser. Par la suite, elle a été enrichie pour prendre en compte le contexte (situation) d'utilisation des processus (Ben Said et al., 2018). Quant à nous, cette dimesnion décrit un SBP dans son ensemble, à travers la spécification de ses principales caractéristiques et les différents types d'informations contextuelles impliqués dans ce processus (Rosemann et al., 2008), tels que, les différentes typologies de BP/SBP (*e.g.,* les processus inter-organisationnels, les processus stratégiques, etc.), les intentions distales qui planifient, contrôlent et réalisent les actions de SBP, les objectifs à atteindre, le contexte d'utilisation de SBP, les résultats fournis, les clients, etc. Ces éléments contextuels permettent de traiter et d'assurer la flexibilité des SBP (Ben Hassen et al., 2018).

3.1.6. La dimension connaissance

La dimension connaissance vient enrichir la description multi-dimensionnelle des BP par une nouvelle dimension décrivant tous les aspects de KM. Cette dimension étendue

représente la dimension centrale et la plus pertinente pour la caractérisation des SBP. Concrètement, elle se focalise sur les flux de connaissances et la dynamique d'acquisition, de préservation, de conversion, de diffusion, de partage, de développement, et de (ré) utilisation des connaissances individuelles et collectives au sein et entre les activités de SBP. En outre, elle considère les différents types de connaissances (*e.g.,* l'aspect crucialité, l'aspect tacite/explicite, l'aspect déclarative/procédurale, l'aspect interne/externe des connaissances, etc.) mobilisées et créés par chaque type d'activité liée au SBP, les différentes sources de connaissances, leur localisation et les endroits où elles sont utilisables ou utilisées, leur nature, ainsi que leur mode de diffusion (*i.e.,* l'aspect individuelle/collective des connaissances) (Ben Hassen et al., 2017b ; 2017c).

En somme, ces six dimensions de BPM-KM sont complémentaires et sont indispensables pour une conceptualisation et une représentation complète et expressive des SBP, afin d'améliorer l'identification et la gestion des connaissances (Ben Hassen et al., 2017a ; 2017 b ; 2017c ; 2018). Dans la suite, nous présentons les nouvelles exigences de modélisation des SBP qui sont relatifs à ces différentes dimensions et qui sont importantes pour la représentation à la fois des aspects statiques et des aspects dynamiques de SBP.

3.2. Exigences de modélisation des processus métier sensibles

À mesure que les modèles de SBP deviennent plus complexes, un langage approprié de BPM devrait satisfaire les exigences sous-jacentes de modélisation des SBP en couvrant parfaitement tous ses aspects pertinents, *i.e.,* les dimensions fonctionnelle, organisationnelle, connaissance informationnelle, comportementale et intentionnelle). Il convient de préciser que chaque exigence donnée doit satisfaire une (ou plusieurs) dimension(s). Par exemple, les exigences E1, E2, E3 et E4 sont attachées à la dimension fonctionnelle. En revanche, certaines exigences peuvent être liées à plusieurs dimensions. Par exemple, E3 qui répond à la modélisation des caractéristiques de la dimension fonctionnelle, peut satisfaire ainsi les caractéristiques de la dimension connaissance.

1- Exigences relatives à la modélisation de la dimension fonctionnelle de SBP

− **E1− Modélisation de la dimension individuelle/collective des actions** : un langage approprié pour la modélisation des SBP devrait fournir des concepts permettant de différencier les actions individuelles des actions collectives (*e.g.,* les actions d'organisation, les actions d'unités organisationnelles, les actions inter-organisationnelles, etc.) lors de la modélisation des SBP, avec des définitions non ambiguës des concepts. Ces actions peuvent être des actions préméditées/délibérées. La prise en compte de la perspective individuelle/collective des actions est très importante dans notre contexte d'étude, du fait que nous nous intéressons à la localisation et l'identification des connaissances nécessaires pour la réalisation des SBP. Ces connaissances prises dans l'action peuvent être, soit individuelles, soit collectives/organisationnelles (tacites ou explicites).

− **E2− Modélisation de la dimension critique et de la dimension intensité de connaissance des actions/des activités organisationnelles**: un langage pour la modélisation des SBP devrait fournir des concepts permettant de modéliser la perspective critique ainsi que la perspective intensité de connaissances des activités organisationnelles (individuelles et collectives), qui sont nécessaires pour déterminer les connaissances cruciales qui sont mobilisées et créées par ces activités, dans une perspective d'améliorer l'échange, la diffusion et la génération des connaissances.

− **E3– Modélisation des différentes opportunités de conversion et de création des connaissances** : un langage pour la modélisation des SBP devrait fournir des concepts

permettant, d'une part, de spécifier les différentes opportunités de conversion et de transfert des connaissances (*i.e.,* la socialisation, l'internalisation, l'explicitation, l'externalisation et la combinaison), et d'autre part, de représenter les flux de conversion et de transfert des connaissances entre les différents types de connaissances (*e.g.,* les connaissances tacites, les connaissances explicitées, les connaissances individuelles et les connaissances collectives) et les activités de conversion de ces connaissances. Par exemple, l'action *Socialisation* qui convertit et transmet des connaissances tacites individuelles en des (nouvelles) connaissances tacites collectives, à travers les interactions informelles. En outre, ce langage devrait être capable de spécifier plus de deux possibilités de conversions de connaissances qui se déroulent dans une seule activité composant un SBP.

− **E4 – Modélisation des interactions collaboratives** : un langage pour la modélisation des SBP devrait fournir des concepts permettant de modéliser des activités collaboratives et/ou les actions inter-organisationnelles impliquant la coopération de plusieurs entités agentives qui sont nécessaires pour la réalisation des SBP interactifs. Ainsi, ce langage devrait spécifier interactions (inter)humaines réelles qui se produisent dans les SBP, au cours desquelles les agents interagissent, échangent, partagent des informations et des connaissances (à travers des messages), et en génèrent de nouvelles.

2- Exigences relatives à la modélisation de la dimension organisationnelle de SBP

− **E5− Modélisation des différents rôles des agents (qui détiennent des connaissances)** : un langage pour la modélisation des SBP devrait fournir des concepts permettant de prendre en compte les rôles des agents opérants dans les activités de SBP, qui peuvent être, par exemple, des individus, des groupes/ des collectifs ou des organisations, etc. qui interagissent, communiquent, échangent, partagent et créent des connaissances (tacites) et /ou des informations (sources de connaissances).

3- Exigences relatives à la modélisation de la dimension connaissance de SBP

− **E6 − Modélisation des connaissances mobilisées et produites par les activités organisationnelles** : un langage pour la modélisation des SBP devrait distinguer les connaissances utilisées pour réaliser un BP/SBP (*knowledge input*) des connaissances créées comme résultat de l'exécution des différents types d'activités organisationnelles (individuelles et collectives) de SBP (*knowledge output*).

− **E7 − Localisation des connaissances** : un langage pour la modélisation des SBP devrait fournir des concepts permettant de représenter les différentes sources de connaissances, *e.g.,* des individus, des experts, des collectives, des supports physiques de connaissances (*e.g.,* des documents, un système à base de connaissances, une mémoire organisationnelle, etc.) qui sont utilisées, générées et/ou modifiées par les activités de SBP, ainsi que leur localisation (où les connaissances peuvent être acquises et clairement explicitées (*e.g.,* les experts qui détiennent les connaissances tacites).

− **E8 – Modélisation des différents types/natures de connaissances (l'aspect tacite/explicite, l'aspect individuelle/collective, l'aspect crucialité, etc.)** : un langage pour la modélisation des SBP devrait fournir des concepts permettant de distinguer explicitement et représenter séparément les différents types de connaissance qui sont mobilisées et créées par les différents types activités de SBP (*e.g.,* les actions individuelles, les actions collectives, les activités à forte intensité de connaissances, les activités critiques, etc.) : les connaissances tacites, les connaissances explicites, les connaissances explicitables, les connaissances individuelles, les connaissances collectives, les connaissances stratégiques, les connaissances procédurales, les connaissances externes, etc.). Cette distinction permet, d'une part, de localiser les connaissances et, d'autre part, de spécifier les différentes potentialités de conversion des connaissances.

4- Exigences relatives à la modélisation de la dimension informationnelle de SBP

− **E9− Modélisation des différents types de ressources** : un langage pour la modélisation des SBP devrait fournir des concepts pour représenter les ressources (*i.e.,* les ressources matérielles, les ressources immatérielles et les ressources humaines) qui permettent de réaliser des actions de SBP.

− **E10 – Différenciation et modélisation des données et des informations des connaissances** : un langage pour la modélisation des SBP devrait fournir des concepts permettant de représenter et séparer les données et les informations des connaissances dans la représentation des flux entre les activités de SBP (comme des inputs et/ou des outputs). En effet, l'échange d'informations et de données (dans le contexte d'une interaction entre les agents) constitue la base de partage et de génération de nouvelles connaissances.

− **E11− Modélisation et différenciation entre les sources de données, d'informations et de connaissances** : un langage pour la modélisation des SBP devrait fournir des concepts permettant de présenter et de séparer les sources de données, d'informations et de connaissances qui sont nécessaires à la réalisation des activités de SBP de celles qui sont générées, créées comme résultats de la réalisation des activités.

5- Exigences relatives à la modélisation de la dimension comportementale de SBP

− **E12 − Modélisation des flux de données, d'informations et de connaissances** : un langage pour la modélisation des SBP devrait fournir des concepts permettant d'illustrer et de représenter les flux de données, les flux d'informations et les flux de connaissances entre les différentes sources (de données, d'informations et de connaissances) et les activités.

− **E13 – Modélisation des processus flexibles hautement dynamiques** : un langage pour la modélisation des SBP devrait modéliser des BP complexes hautement dynamique qui peuvent être non structurés ou semi-structurés, nécessitant une flexibilité substantielle, une coordination et une collaboration entre les parties prenantes du BP. Ainsi, ce langage devrait traiter les exceptions imprévues qui se produisent durant leur exécution, ainsi que les règles et les contraintes qui influencent la structure du processus.

6- Exigences relatives à la modélisation de la dimension intentionnelle de SBP

− **E14– Modélisation des intentions et des objectifs** : un langage pour la modélisation des SBP devrait fournir des concepts permettant de modéliser les intentions distales individuelles et collectives. Ces éléments sont considérés comme une justification pour la réalisation des actions individuelles et collectives, pour l'atteinte des objectifs, pour la prise de décision guidant le flux ad hoc de SBP, la séquence d'activités à chaque instance du processus, ainsi que pour la collaboration et l'échange de connaissances.

Dans ce qui suit, ces exigences sont utilisées pour analyser et évaluer la capacité représentative d'un ensemble sélectionné d'approches et de langages, fréquemment étudiés dans la littérature des domaines de BPM-KM, pour supporter parfaitement un SBP.

4. Approches de modélisation des processus métier sensibles

La modélisation des SBP est une opération complexe. Malgré le fait qu'il existe dans la littérature une multitude d'approches et de langages de BPM pour décrire les BP, les différentes catégories proposées ne répondent pas à nos besoins de modélisation des SPBs. Ainsi, nous avons proposé deux grandes classes d'approches et de langages susceptibles de modéliser ce type particulier de BP (qui ne possèdent pas le même niveau d'expressivité).

4.1. Approches et langages de modélisation des BP/Workflows conventionnels

Certains langages conventionnels/traditionnels de BPM ont été adaptés pour supporter implicitement la représentation des éléments intrinsèques de connaissances dans les modèles de BP. Nous citons comme exemple : eEPC (Extended Event Driven Process Chain) (Wagner and Klueckmann, 2006), (Sheer, 2013), UML 2.0 AD (UML 2.0 Activity Diagrams) (OMG, 2011), BPMN 2.0.2 (Business Process Modeling Notation) (OMG, 2013), CMMN (Case Management Model and Notation) (OMG, 2016a), DMN 1.1 (Decision Model and Notation) (OMG, 2016b), etc. Cette classification de langages (évolués et à usage général) est largement utilisée et adoptée dans les scénarios pratiques au sein des organisations. Ces langages sont appropriés pour la modélisation de la perspective de processus dans son ensemble. La plupart d'entre eux se concentrent davantage sur la représentation de processus « déterministes », ayant une faible complexité.

4.2. Approches et langages de modélisation des connaissances orientés processus

L'intégration des BPs et des flux de connaissances a retenu l'attention des communautés de recherche qui est rapidement devenue un sujet de recherche très intéressant. Dans ce contexte, la littérature présente un ensemble d'approches et de notations pour modéliser les aspects de connaissance/de KM dans les modèles de BPs, particulièrement, les processus à forte intensité de connaissances (KIP) (Gronau et al., 2005), (Di Ciccio et al., 2015). Cette classification d'approches et de langages de modélisation des connaissances orientés processus inclut : l'approche DECOR (Abecker, 2001), l'approche CommonKADS (Schreiber et al., 2002), la méthode BPKM (Business Process Knowledge Method) proposée par (Papavassiliou and Mentzas, 2003), l'approche KTA (Knowledge Transfer Agent) (Strohmaier et al., 2007), PROMOTE (Process-oriented methods and tools for knowledge management) (Woitsch and Karagiannis 2005), GPO-WM (Heisig 2006), KMDL (Knowledge Modeling Description Language) ((Gronau et al., 2005), (Arbeitsbericht, 2009)), MailofMine (Di Ciccio et al., 2015), DCR Graphs (Hildebrandt and Mukkamala, 2010), KIPN (Knowledge-intensive Process Notation) (Netto et al., 2013), etc. La majorité d'entre eux se focalisent sur le stockage et le transfert des connaissances. Cependant, ils incluent partiellement la perspective de processus dans son ensemble.

Une description détaillée des différentes classifications des approches et des langages de modélisation proposées est présentée dans (Ben Hassen et al., 2017b ; 2018).

4.3. Analyse des approches et des langages selon les exigences de modélisation des SBP

Le Tableau 2 présenté ci-après illustre une analyse comparative des approches et des notations de modélisation présentées concernant leur expressivité et leur capacité à représenter parfaitement et explicitement les exigences spécifiques relatives aux différentes dimensions de modélisation des SBP proposées précédemment.

Tout compte fait des résultats de l'évaluation, les approches et les langages de modélisation existants ne sont pas parfaitement appropriés pour couvrir adéquatement et complètement les particularités et les exigences de modélisation de SBP citées précédemment (Cf. E1− E14). Les dimensions fonctionnelle, organisationnelle, comportementale, informationnelle et intentionnelle sont mieux supportées dans la catégorie des approches et des langages de BPM conventionnels (UML AD, eEPC, BPMN, CMMN), mais avec des capacités de représentation plus ou moins limités. En revanche, ils sont extrêmement limités pour la modélisation de la dimension connaissance (E6, E7 et E8). Par exemple, ces différents langages ne distinguent pas clairement et explicitement

entre les définitions des différents types activités, y compris, entre autres, les actions individuelles, les actions collectives (*e.g.,* les actions d'organisation, les actions d'unité organisationnelle, les actions collaboratives, les actions inter-organisationnelles, etc.) (**E1** −/+] ainsi que les activités critiques et les activités à forte intensité de connaissances [**E2** −]. De même, ils ont des capacités limitées pour modéliser explicitement les aspects d'interactions (inter)humaines [**E3** −]. Ainsi, ils ne conviennent pas pour supporter adéquatement les aspects dynamiques relatifs à la flexibilité (des changements prévisibles et non prévisibles dans les activités de SBP) [**E1**3 −/+], (excepté BPMN et CMMN).

Tableau 2. Évaluation des approches et des langages de modélisation selon les différentes exigences de modélisation des SBP ("-" signifie « l'approche/le langage ne supporte pas une exigence donnée »; "- / +" signifie « supporte partiellement une exigence donnée » ; "+" signifie « supporte une exigence donnée »)

	Approches et langages de BPM /Workflow conventionnels				**Approches et langages de modélisation de connaissances orientées processus**									
Exigences de modélisation des SBP/ dimension	**UML 2.0 AD**	**eEPC**	**BPMN 2.0.2**	**CMMN 1.1**	**Common KADS**	**BPKM**	**DECOR**	**GPO-WM**	**KTA**	**KMDL**	**PROMOTE**	**KIPN**	**DCR Graphs**	**MailofMine**
Dimension fonctionnelle														
E1	-/+	-/+	-/+	-/+	-	-/+	-	-	-	-/+	-/+	-	-	-
E2	-	-	-	-	-	-/+	-	-/+	-	-/+	-/+	+	-	-
E3	-	-	-	-	-	-	-	-	-	+	+	-/+	-	-
E4	-	-/+	+	-	-	-	-	-	+	-/+	-	+	-/+	-/+
Dimension organisationnelle														
E5	-/+	-/+	+	-	-/+	-/+	-/+	-	+	-/+	-/+	-/+	-	-
Dimension connaissance														
E6	-	-/+	-	-	-/+	-/+	-/+	-	+	+	-	-	-	-
E7	-	-/+	-/+	-	-	+	-	+	-/+	-/+	-/+	-/+	-	-
E8	-	-/+	-	-	-	-	-	-/+	-/+	-/+	-/+	-/+	-	-
Dimension informationnelle														
E9	-/+	-/+	+	-	-/+	-	-/+	-	+	-/+	-/+	-	-	-
E10	-/+	+	+	-/+	-/+	-/+	-/+	-/+	-	-/+	-/+	-	-	-
E11	-/+	-/+	-/+	-/+	-	-/+	-	-	-	-	-/+	-/+	-/+	-/+
Dimension comportementale														
E12	-/+	-/+	-/+	-	-	-	-	-	-/+	-/+	-/+	-	-	-
E13	-/+	-/+	+	+	-	-/+	-	-	-	-	-	+	-/+	+
Dimension intentionnelle														
E14	-	-/+	-	-/+	-	-	-	-	+	-	-	+	-	-

En revanche, BPM est un défi pour les langages de modélisation des connaissances (PROMOTE, KMDL, KIPN) qui se concentrent davantage sur la conversion des connaissances [**E3** +]. Ces notations ont des capacités limitées pour modéliser complétement et adéquatement la perspective de processus et la logique /les flux de contrôle de SBP, si on les compare à BPMN et ARIS eEPC. Dans le même temps, elles sont inappropriées pour représenter parfaitement et complétement les exigences et les

aspects pertinents relatifs à la dimension connaissance (*e.g.,* la distinction explicite entre les données, les informations et les connaissances qui sont mobilisées et produites par les différents types d'activités [**E10** −/+], les différents types de connaissances (*e.g.,* l'aspect individuelle/collective, l'aspect tacite/explicite, l'aspect factuelle/procédurale des connaissances) [**E8** −/+], les différentes sources de connaissances [**E7/E11** −/+] et la différenciation entre les flux de connaissance et les flux d'information [**E12** −/+]). Ces notations ne fournissent pas de symboles spécifiques pour leur représentation (séparément) dans les modèles de BP. Le problème réside principalement dans l'absence d'une compréhension claire et commune de la relation entre ces différentes notions. L'incertitude n'existe pas seulement dans les différentes définitions, mais également dans l'utilisation pratique de ces concepts dans les modèles de BP. Par conséquent, ces différentes lacunes mènent à développer des modèles de SBP incomplets, ambigus et incompréhensibles.

Dans les perspectives de pallier aux différentes limitations détectées et répondre adéquatement aux nouvelles exigences de modélisation des SBP, la spécification d'une conceptualisation précise, rigoureuse, commune et consensuelle de SBP avec une notation de modélisation appropriée qui intègre tous les enjeux/aspects appropriés au couplage de BPM-KM dans les modèles de SBP, sont d'une importance primordiale. Un tel langage devrait supporter et intégrer explicitement et adéquatement toutes les perspectives de modélisation pertinentes de SBP, *i.e.,* les dimensions fonctionnelle, organisationnelle, comportementale, informationnelle, intentionnelle et connaissance, afin de développer des modèles de représentation graphiques complets et expressifs.

6. Conclusion

Ce papier introduit la problématique de l'analyse conceptuelle des SBPs dans une perspective d'identification et de gestion des connaissances (cruciales). Ces processus sont fortement complexes et à haute intensité de connaissances. L'originalité de cette contribution repose sur l'approche multi-dimensionnelle que nous avons adopté pour la modélisation des SBP. En premier lieu, nous avons proposé une caractérisation rigoureuse pour ce type de processus (qui le distinguent des BP classiques). En second lieu, nous avons défini six dimensions de caractérisation des SBP (*i.e.,* les dimensions fonctionnelle, organisationnelle, comportementale, informationnelle, intentionnelle et connaissance) ainsi que des exigences spécifiques pour leur modélisation qui sont relatives à ses différentes dimensions. En dernier lieu, nous avons mené une analyse comparative des différentes approches et langages de modélisation actuels (certains sont centrés BP d'autres sont centrés connaissances) selon les différentes exigences proposées pour en déduire leur expressivité et leur capacité à représenter parfaitement et explicitement les particularités de SBP. Tout compte fait des résultats de l'évaluation, aucune des approches et de notations existantes ne satisfait, individuellement, toutes les caractéristiques et les nouvelles exigences de modélisation des SBP. De la liste des langages sélectionnés, le standard BPMN 2.0.2 (OMG, 2013) semble être la notation la plus prometteuse pour enrichir la modélisation des SBP qui couvre la plupart de ses dimensions (orientées modélisation métier), quoiqu'elle soit trop faible dans la modélisation de la dimension connaissance.

L'analyse conceptuelle des SBP présente plusieurs apports dans la continuité de nos travaux. Nos activités de recherche actuelles se focalisent principalement sur deux aspects pertinents. Le premier aspect consiste à définir une spécification conceptuelle formelle, cohérente et rigoureuse des différentes dimensions de caractérisation de SBP, sous forme d'une ontologie noyau, que nous appelons COSBP (Core Ontology of Sensitive Business Processes). Cette ontologie spécialise l'ontologie fondatrice DOLCE « Descriptive

Ontology for Linguistic and Cognitive Engineering » (Masolo et al., 2003), et étend et complète l'ontologie noyau des processus d'organisation COOP (Core ontology of Organization's Processes) (Turki et al., 2016). COSBP offre un référentiel de concepts (et de relations sémantiques) génériques et consensuels de SBP classés par catégorie de six classes de modules ontologiques relatifs aux six dimensions de modélisation des SBP, en se basant sur des définitions explicites et sémantiquement riches des concepts mis en jeu. Le deuxième aspect consiste à développer une extension du langage BPMN 2.0.2 incluant une notation graphique pour les SBP dans une perspective de gestion des connaissances. Soulignons que, des tentatives d'extension ont été déjà proposées dans des travaux de recherche antérieurs (Ben Hassen et al., 2017a ; 2017c) en intégrant certains concepts clés de SBP relatifs à la dimension fonctionnelle et la dimension connaissance. La nouvelle extension « BPMN4SBP » permet de remédier aux différentes lacunes recensées et supporter parfaitement la modélisation multi-perspective des SBP, en intégrant et implémentant tous les concepts définis dans les modules ontologiques de COSBP.

Bibliographie

Abecker, A. (2001). DECOR Consortium : DECOR—Delivery of Context- Sensitive Organizational Knowledge, E-Work and E-Commerce. IOS Press, Amsterdam

Arbeitsbericht, (umfangreicheBeschreibung) (2009) KMDL®v2.2. http://www.kmdl.de

Ben Hassen, M., Turki M., Gargouri, F. (2017a). Extending sensitive business process modeling with functional dimension for knowledge identification. In Proceedings of the 14th International Conference on e-Business (ICE-B 2017), Madrid, Spain., Vol. 2, pp. 38-51, SciTePress.

Ben Hassen, M., Turki M., Gargouri, F. (2017b). Towards Extending Business Process Modeling Formalisms with Information and Knowledge Dimensions. In Proceedings of the 30th International Conference on Industrial and EngineeringApplications of Artificial Intelligence and Expert Systems (IEA/AIE' 2017), Arras, France., Vol. 10350, Springer 2017.

Ben Hassen, M., Turki M., Gargouri, F. (2017c). Using core ontologies for extending sensitive business process modeling with the knowledge perspective. In Proceedings of the Fifth European Conference on the Engineering of Computer-Based Systems (ECBS'2017), Cyprus (p.2). ACM.

Ben Hassen, M., Turki M., Gargouri, F. (2018). Comparative Analysis of Contemporary Modeling Languages Based on BPM4KI Meta-Model for Sensitive Business Processes Representation. International Journal of Enterprise Information Systems (IJEIS), 14(3), pp.41-78, 2018

Ben Said, I., Chaabane, M., Bouaziz, R., & Andonoff, E. (2018). BPMN4VC-modeller: easy-handling of versions of collaborative processes using adaptation patterns. International Journal of Information Systems and Change Management, 10(2), 140-189.

Bušinska, L. and Kirikova, M. (2011). Knowledge Dimension in Business Process Modeling. In Information Systems in a Diverse World: Selected Extended Papers at CAiSE Forum. London: Springer, 186-201.

Di Ciccio, C., Marrella, A., & Russo, A. (2015). Knowledge-intensive processes: characteristics, requirements and analysis of contemporary approaches. Journal on Data Semantics, 4(1), 29-57.

Ghrab, S., and Saad, I. (2016). Identifying Crucial Know-How and Knowing-That for Medical Decision Support. IJ of Decision Support System Technology (IJDSST), 8(4), 14-33.

Gronau, N., Korf, R. and Müller, C. (2005). KMDL Capturing, Analysing and Improving Knowledge-Intensive Business Processes. Journal of Universal Computer Science, vol. 11, no. 4, pp. 452-472.

Grundstein, M. (2009). GAMETH®: a constructivist and learning approach to identify and locate crucial knowledge. International Journal of Knowledge and Learning, 5(3-4):289–305

Heidari, F., Loucopoulos, P., Brazier, F., &Barjis, J. (2013). A meta-meta-model for seven business process modeling languages. In 15th Conference on Business Informatics (pp. 216-221). IEEE.

Heisig, P. (2006). The GPO-WM® method for the integration of knowledge management into business processes. In: International Conference on Knowledge Management, Graz, pp. 331–337.

Hildebrandt TT, Mukkamala RR (2010) Declarative event-based workflow as distributed dynamic condition response graphs. In: Programming languages approaches to concurrency and communication-cEntric software. Cyprus, pp 59–73

Masolo, C., Borgo, S., Gangemi, A., Guarino, N., Oltramari, A., & Schneider, L. (2003). *The WonderWeb Library of Foundational Ontologies and the DOLCE ontology*". Wonder Web Deliverable D18, Final Report (version 1.0, 31-12-2003).

Netto, J.M, Franca, J. B. S., Baião, F.A. and Santoro, F. M. (2013). A notation for Knowledge-Intensive Processes. Proceeding of the 2013 IEEE 17thInternational Conference on Computer Supported Cooperative Work in Design (CSCWD), 190-195.

OMG, 2011. UML-Unified Modeling Language (2011) V2.4.1. Object Management Group. http://www.omg.org/spec/UML/2.4.1/Superstructure/PDF

OMG, 2013. Business Process Modeling and Notation (BPMN). Version 2.0.2, 2013. http://www.omg.org/spec/BPMN/2.0.2/pdf

OMG, 2016. Case Management Model and Notation (CMMN). Version 1.1. http://www.omg.org/spec/CMMN/1.1

Ouali, S., Mhiri, M., & Bouzguenda, L. (2016). A multidimensional knowledge model for business process modeling. Procedia Computer Science, 96, 654-663.

Papavassiliou, G., Mentzas, G.: Knowledge modelling in weakly-structured business processes. J. Know. Manag. 7(2), 18–33 (2003)

Reichert, M., & Weber, B. (2012). *Enabling flexibility in process-aware information systems: challenges, methods, technologies*. Springer Science & Business Media.

Rosemann, M., Christian, R. Jan, F. (2008). Contextualisation of business processes. International Journal of Business Process Integration and Management, 3(1), pp.47-60.

Saad, I., Grundstein M., Sabroux, C. (2009). Une méthode d'aide à l'identification des connaissances cruciales pour l'entreprise. Revue Systèmes d'Information et Management (SIM), 14 (3),43–78.

Saidani, O., Nurcan, S.(2009). Context-Awareness for Adequate Business Process Modeling. In the International Conference on Research Challenges in Information Science (RCIS), 177-186.

Schreiber, A. T., Schreiber, G., Akkermans, H., Anjewierden, A., Hoog, R., ... & Wielinga, B.(2000). Knowledge engineering and management: the CommonKADS methodology. MIT press.

Turki, M, Saad, I ,Gargouri, F, & Kassel, G. (2014). A Business Process Evaluation Methodology for Knowledge Management based on multi-criteria decision-making approach. In Saad. I, Sabroux. CR, Gargouri. F. (eds.), Information systems for knowledge management. Wiley-ISTE

Turki M, Kassel G, Saad I, Gargouri F (2016) A core ontology of business processes based on DOLCE.J DataSemant5(3) :165–177.

Van der Aalst, W.M.P., Weske, M., Wirtz, G. (2003). Advanced Topics in Workflow Management : Issues, Requirements, and Solutions, Journal of Integrated Design and Process Science, vol. 7, n° 3.

Wagner K., & Klueckmann, J (2006). Business Process Design as the Basis for Compliance Management, Enterprise Architecture and Business Rules. In AGILITY by ARIS Business Process Management (pp. 117–127). Springer, Berlin, Heidelberg.

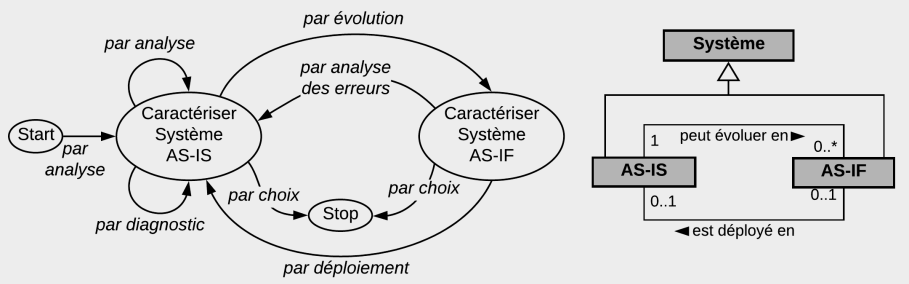
Woitsch R. and Karagiannis, D. (2005). Process Oriented Knowledge Management : A Service Based Approach. Journal of universal computer science 11(4), 565-588.

Un cadre méthodologique As-Is/As-If pour guider le développement des méthodes d'évolution continue - Résumé

Ornela Cela, Mario Cortés-Cornax, Agnès Front et Dominique Rieu

Université Grenoble Alpes, CNRS, Inria, Grenoble INP, LIG, 38000 Grenoble, France prenom.nom@univ-grenoble-alpes.fr

Cet article est un résumé de l'article publié à CAiSE 2019 (Cela *et al.*, 2019) qui propose le cadre méthodologique *As-Is / As-If* permettant d'aider les ingénieurs de méthodes à construire ou à adapter des méthodes d'évolution continue (Singh et Singh, 2015) dédiées à des enjeux d'évolution dans un contexte organisationnel : par exemple l'amélioration des processus métier au sein de l'organisation, celle de la capacité d'absorption des connaissances dans les projets d'innovation de réseaux de PME, l'innovation organisationnelle dans des écosystèmes, etc. Le cadre *As-Is / As-If* est construit selon une approche *bottom-up* qui généralise les concepts et phases proposés dans plusieurs méthodes d'évolution continue développées au cours de projets de recherche. Il propose un modèle de processus formalisé sous la forme d'une MAP intentionnelle (Rolland, 2007) et un méta-modèle de produit formalisé en UML. Le méta-modèle de produit et le modèle de processus peuvent être réutilisés et adaptés pour guider un ingénieur de méthodes dans la construction ou l'adaptation de méthodes d'évolution continue. Ils peuvent être considérés comme des exemples prototypiques à adapter à la situation actuelle à l'aide d'heuristiques, pour concevoir une nouvelle méthode appelée méthode cible. La figure suivante présente la MAP de niveau supérieur représentant la vue générale du modèle de processus et les méta-classes de base du méta-modèle de produit.



Le modèle de processus a deux intentions principales : caractériser le système As-Is et imaginer le système As-If possible, le terme système pouvant désigner un système d'information, un écosystème, un processus métier... Chaque cycle d'évolution consiste tout d'abord à analyser le système actuel (stratégie d'analyse) pour déterminer ses principaux composants susceptibles d'évoluer, et à réaliser un diagnostic (stratégie de diagnostic) de ce système actuel pour déterminer ses points de blocage et proposer des buts à atteindre pour résoudre ces points de blocage. Puis le but est (stratégie d'évolution) d'imaginer des scénarios d'évolution possibles en déterminant un ensemble de systèmes As-If possibles, chacun d'eux correspondant à un scénario d'évolution. L'un de ces multiples systèmes As-If sera alors choisi et consolidé en tenant compte des différentes contraintes juridiques, économiques, sociales, techniques, etc. Si aucun scénario d'évolution n'est choisi en raison de l'impossibilité d'identifier une évolution satisfaisante (stratégie d'analyse de défaillance), l'analyse et le diagnostic pourraient être rejoués. Si un scénario d'évolution est choisi, il sera déployé devenant le nouveau système étudié (stratégie de déploiement). La décision de mettre fin au cycle d'évolution continue doit être un choix collectif de tous les acteurs (stratégie « Par choix »). Le méta-modèle produit (partie droite de la figure) montre une version simplifiée des associations entre les systèmes As-Is et As-If. Il représente les concepts utilisés par le modèle de processus. Les deux modèles illustrent les modèles de haut niveau fournis par le cadre qui servent à raisonner sur le cycle d'évolution et à aider à la prise de décisions lors de leur adaptation à la nouvelle situation en cours. Comme cela a été mentionné auparavant, ce cadre généralise nos expériences empiriques dans la construction des méthodes d'évolution continue ADInnov (Cortes-Cornax *et al.,* 2016), ISEACAP (Movahedian *et al.,* 2017) et CEFOP (Cela *et al.,* 2017). L'article s'appuie sur les CEFOP et ADInnov, pour illustrer l'utilisation du cadre As-Is / As-If.

Bibliographie

Cortes-Cornax, M., Front, A., Rieu, D., Verdier, C., Forest, F.: ADInnov: An Intentional Method to Instil Innovation in Socio-Technical Ecosystems. In: International Conference on Advance Information System Engineering (CAiSE 2016), Springer, pp. 133–148

Movahedian, F., Front, A., Rieu, D., Farastier, A. et al.: A participative method for knowledge elicitation in collaborative innovation projects. In : International Conference on Research Challenges in Information Science (RCIS 2017), IEEE, pp. 244–254

Cela, O., Front, A., Rieu, D.: CEFOP: A method for the Continual Evolution of Organisational Processes. International Conference on Research Challenges in Information Science (RCIS 2017), IEEE, pp. 33–43

Ornela Cela, Mario Cortes Cornax, Agnès Front, Dominique Rieu: Methodological Framework to Guide the Development of Continual Evolution Methods. CAiSE 2019: 48-63

Rolland,C.: Capturing System Intentionality with Maps. In: Conceptual Modelling Information System Engineering, Springer-Verlag (2007), pp. 141–158

Singh, J. Singh, H.: Continuous improvement philosophy– literature review and directions. IN: Benchmarking: An International Journal (2015), vol. 22 (1), pp. 75-119

Tiers-Lieu pour les services d'information

La valeur de la modélisation conceptuelle

Jolita Ralyté, Michel Léonard

ISS, CUI, Université de Genève
Battelle bâtiment A, 7 Route de Drize, 1227 Carouge, Suisse
jolita.ralyte@unige.ch, michel.leonard@unige.ch

RÉSUMÉ. Il s'agit ici d'un résumé étendu de notre article (Ralyté, Léonard, 2019). Pour réussir, la transformation numérique ne peut pas être considérée comme relevant d'un seul département ou d'une seule organisation. Sa mission consiste à fournir de nouveaux services d'information interdisciplinaires voire transdisciplinaires. L'intelligence collective est donc la clé de son succès. Une approche collaborative facilitant l'innovation et la cocréation est nécessaire pour développer des services d'information durables et responsables. En tant que solution potentielle, nous proposons une nouvelle approche appelée Tiers-Lieu pour les Services (TLS). Dans cette approche, nous considérons la modélisation conceptuelle comme la technique centrale et fondamentale pour la coconception de services d'information et donc pour la réussite d'un TLS. Nous présentons également deux TLS expérimentaux.

ABSTRACT. This is an extended summary of our article (Ralyté, Léonard, 2019). To be successful, Digital Transformation cannot be considered as a matter of a single department or a single organization. Its mission consists in providing new inter-disciplinary or even transdisciplinary information services. Therefore, collective intelligence is a key for its success. A collaborative approach enabling innovation and co-creation is necessary to develop sustainable and responsible information services. As a potential solution, we propose a novel approach called Tiers-Lieu for Services (TLS). In this approach, we consider conceptual modeling as the central underpinning technique for the co-design of information services and therefore for the success of a TLS. We also report on two experimental TLS.

MOTS-CLÉS : Tiers-Lieu pour les services, transformation digitale, modélisation conceptuelle, service d'information, service transdisciplinaire

KEYWORDS: Tiers-Lieu for Services, Digital Transformation, Conceptual Modeling, Information Service, Transdisciplinary Service

1. Contexte et objectifs

À l'ère de la transformation numérique, les organisations, privées comme publiques, cherchent constamment à innover dans leur façon de développer et de fournir des services et des produits, à être plus performantes et attractives, et à transformer leurs modèles d'activité grâce aux technologies numériques. Par

ailleurs, cette transformation n'est pas uniquement axée sur la technologie. Elle embrasse des ambitions stratégiques beaucoup plus larges et implique des changements fondamentaux dans les activités, la structure et même la culture des organisations (Baker 2014). On peut même dire que le progrès de la société dépend en grande partie du succès de sa transformation numérique. L'enjeu consiste à construire une infrastructure numérique prenant la forme de services d'information et des systèmes de services interdisciplinaires voire transdisciplinaires. Compte tenu de la variabilité des activités, des situations et des intentions à prendre en considération pour construire cette infrastructure numérique, il est essentiel que tous les acteurs concernés contribuent. Une approche contributive et exploratoire soutenant l'innovation et la cocréation est nécessaire pour développer des services d'information responsables et durables. A cet effet, nous proposons une nouvelle approche appelée Tiers-Lieu pour les Services (TLS) (Ralyté et Léonard, 2019, 2020). TLS fournit des moyens pour conduire la contribution de plusieurs parties prenantes : faire émerger et explorer les idées, stimuler la créativité, d'éliminer les barrières cognitives et sociales, et, le plus important, de cocréer des services d'information transdisciplinaires. Nous affirmons également que la modélisation conceptuelle est la clé pour rendre la cocréation des services d'information efficace.

2. Tiers-Lieu pour les services et son exploration

Selon Burret (2017), un Tiers-Lieu est « une configuration sociale où la rencontre entre des entités individuées engage intentionnellement à la conception de représentations communes ». Une « représentation commune » est considérée ici au sens large ; il peut s'agir d'une conception d'un service ou d'un système numérique, d'un modèle d'affaires, d'un projet de loi, d'un plan de transport en commun, etc. Une « configuration sociale » signifie que l'intention du Tiers-Lieu touche des personnes d'horizons divers, citoyens responsables ou représentant des organisations publiques ou privées, provenant des domaines d'activité différents.

L'objectif d'un Tiers-Lieu pour les Services (TLS) consiste à offrir un contexte multidisciplinaire et multiinstitutionnel pour la coconstruction des biens communs d'information sous forme des services d'information (Ralyté et al., 2015) interdisciplinaires et interinstitutionnels, voire transdisciplinaires et transinstitutionnels. C'est une approche collaborative impliquant des personnes contributrices hétérogènes, exerçant des métiers et responsabilités différents dans des organisations variées, mais tous visant à innover et à cocréer de la valeur, et donc à contribuer au progrès numérique de la société. Les services d'information construits dans le cadre d'un TLS ne sont pas seulement novateurs, ils sont nécessairement démocratiques, responsables et indispensables.

La conduite d'un TLS consiste en quatre étapes itératives : (1) la définition d'une intention dans un contexte de la progression numérique de la Société, (2) la constitution d'un réseau hétérogène de contributeurs, (3) l'exploration dans le but d'identifier des propositions de valeur et des services d'information, (4) la conception et la mise en place des services d'information. Les sessions d'exploration et de conception utiliseront différentes techniques de cocréation et de

conceptualisation. En effet, la modélisation conceptuelle est la technique sous-jacente à la coconception des services d'information (Ralyté et al., 2015). Le rôle d'un modèle conceptuel est d'assurer une représentation formelle et sans ambiguïté d'un domaine particulier en vue de la numérisation. Ainsi, un TLS doit être piloté par la création, la discussion et le raffinement de modèles conceptuels.

Dans le cadre de notre programme de formation continue, nous avons réalisé deux TLS exploratoires, chacun d'une durée d'environ 15 heures sur deux jours. 13 personnes, étudiants de ce programme, représentant des métiers et des organisations différents, ont participé aux expériences. Le premier TLS était dédié à l'exploration de la poussée technologique – le potentiel de mise en œuvre d'un nouvel artefact numérique et l'identification de services d'information qui pourraient être développés sur la base de cette technologie. Le deuxième TLS, a pris le chemin inverse en explorant un enjeu sociétal nécessitant le développement des services d'information. Les deux explorations ont démontré que le fait d'avoir des compétences en modélisation conceptuelle a un impact direct sur le succès d'un TLS. Par ailleurs, non seulement les techniques conventionnelles de modélisation des données et des processus se sont révélées efficaces, mais aussi les approches de conceptualisation et d'exploration sous forme des modèles d'affaires, de valeurs et de services (ex. Business Model Canvas (Osterwalder et Pigneur, 2010), Service Model Canvas (Turner 2015), Value Proposition Canvas (Osterwalder et al., 2014)) ont joué le rôle de propulseurs d'intelligence collective. Nous avons constaté que la modélisation conceptuelle permet de mieux partager la compréhension du domaine et d'atteindre un consensus informationnel.

Bibliographie

Baker, M. (2014). *Digital Transformation*. CreateSpace Independent Publishing Platform.

Burret, A. (2017). *Étude de la configuration en Tiers-Lieu – la repolitisation par le service*. Thèse de doctorat, Université des Lumières, Lyon, France.

Ralyté, J., Khadraoui, A., Léonard, M. (2015). Designing the Shift from Information Systems to Information Services Systems. Business and Information Systems Engineering, 57(1): 37-49, Springer.

Ralyté, J., Léonard, M. (2019). Exploring the Concept of "Tiers-Lieu" for Information Services: The Value of Conceptual Modeling. ER Forum 2019: 98-107, CEUR-WS 2469,

Ralyté, J., Léonard, M. (2020). Tiers-Lieu for Services: An Exploratory Approach to Societal Progression. In: Nóvoa, M.H, Dragoicea, M. Kühl, N. (eds): Exploring Service Science. IESS 2020, LNBIP 377, Springer, pp. 289-303.

Osterwalder, A., Pigneur, Y. (2010). *Business Model Generation: A Handbook for Visionaries, Game Changers and Challengers*. John Wiley & Sons, Inc.

Turner, N. (2015). Introducing the service model canvas. http://www.uxforthemasses.com/service-model-canvas/

Osterwalder, A., Pigneur, Y., Bernarda, G. (2014). *Value Proposition Design: How to Create Products and Services Customers Want*. John Wiley & Sons, Inc.

Ingénierie logicielle

Practices to Define Software Measurements - *Káthia Marçal de Oliveira* (article long)

A Unified Vision of Configurable Software- *Houssem Chemingui, Inès Gam, Raúl Mazo, Henda Ben Ghezala et Camille Salinesi* (article court)

Modélisation graphique des environnements proxémiques basée sur un DSL - *Paulo Pérez, Philippe Roose, Marc Dalmau, Yudith Cardinale, Nadine Couture et Dominique Masson* (article long)

Xatkit: A model-based chatbot development framework - Extended Abstract - *Gwendal Daniel, Jordi Cabot, Laurent Deruelle et Mustapha Derras* (résumé étendu)

Practices to Define Software Measurements

Káthia Marçal de Oliveira

LAMIH CNRS UMR 8201
Université Polytechnique Hauts-de-France, Valenciennes, France
kathia.oliveira@uphf.fr

RESUME. *Métriques, mesures, indicateurs, estimations, etc. Bien que nommés de différentes manières et explorant différents angles, un fait est reconnu : les mesures de systèmes logiciels sont essentielles pour évaluer leur qualité, favoriser leur amélioration et contrôler leur production. Différentes méthodologies ont été définies (GQM, GQIM, PSM, etc.). Différentes études pratiques ont été publiées. Cependant, la définition des nouvelles mesures semble toujours une tâche non triviale. Dans cet article, nous présentons notre expérience sur la définition des mesures avec un ensemble de pratiques simples qui abordent des questions clés concernant différentes méthodologies. Ces pratiques ont été appliquées dans la définition de mesures pour différents types de systèmes (des « legacy systems » aux applications IoT modernes).*

ABSTRACT. *Metrics, measures, measurements, indicators, estimates and so on. Although named in different ways and exploring different angles, one fact is recognized: measuring software systems is essential for assessing their quality, promoting their improvement and controlling their production. Different methodologies have been defined (GQM, GQIM, PSM, etc.). Different studies in practice have been published. However, defining new measures still seems a not trivial task. In this article, we present our experience on how to define measures with a set of simple practices that address main issues of different methodologies. These practices have been applied in the definition of measures for different types of systems (from legacy systems to modern IoT applications).*

Mots-clés : mesure logicielle, métriques, indicateurs de qualité.

KEYWORDS: *software measurement, measures, metrics, quality indicators.*

1. Introduction

Web systems, Ubiquitous computing, Internet of Things (IoT), Smart Cities, and so on brought a significant diversity of software systems that support several of our daily activities. To ensure the improvement and adoption of these applications, it is essential to assess their quality. Measurements can help address some of the most critical issues in software development and provide support for evaluating, improve, and control the production of software systems (Briand *et al.* 2002). In this context, several well-defined classical measures can be applied (e.g. complexity of the code,

size, coupling, defect density, reuse, etc.). However, these new kinds of applications present specific particularities that need also to be measured to assure their quality. We quote for instance, the need to evaluate context-awareness, a common feature of personalized systems, IoT and ubiquitous software applications, that implies in the perceived quality of the software system (Lee and Yun, 2012; Carvalho *et al.* 2017, 2018). Several other studies to evaluate the quality of different particularities of these new kinds of systems can be found in literature, such as: privacy in ubiquitous systems (Jafari *et al.*, 2011), transparency interaction in smart homes (Wu and Fu, 2012), user immersion degree in ubiquitous services (Lee and Yun, 2012), trust in adaptive systems (Evers *et al.*, 2010)), efficiency of data transfer in IoT applications (Paschou *et al.*, 2013). However, besides not being exhaustive, the defined measures are not always described in a way that allows their use. Carvalho *et al.* (2017) showed that more than 80% of measures found in a systematic review for ubiquitous applications were not formally defined. We are, therefore, faced with defining or redefining measures in order to evaluate these new kinds of systems.

Several works have been presented in the literature for the definition of measures, improvement of process, and institutionalization of measurement programs in the industry (see, for instance, an overview of these approaches in Tahir *et al.*, 2016). Nevertheless, defining new measures is always considered a complex activity. First of all, we need experts for the definition and also for the interpretation of the results, and they are not always available. Once we have experts, we are faced with the lack of adequate tools to collect and evaluate data. Moreover, the definition of the experimental protocols is also complex, there are not always professionals to proceed the evaluations and the evaluation itself is usually expensive (Oliveira *et al*., 2012). Due to these difficulties, it is common to give up measuring and miss good opportunities to perform simple but significant measurement studies.

The idea of this article was to take a step back and organize a set of practices we have applied to deal with those difficulties when defining measures (around of 90) for different software systems (ubiquitous systems (Carvalho *et al.*, 2018), legacy system (Ramos *et al.* 2004), interactive systems (Assila *et al.*, 2016; Gabillon *et al.*, 2013), web systems (Lima *et al.*, 2009)) and software process (Monteiro and Oliveira, 2010). These practices came from the application of different proposals from literature and became a roadmap we have followed to investigate and define measurements. We put together in this paper what worked for us showing different real examples so that it can be applied directly in new definitions of measurement.

We start this paper (section 2) by briefly presenting some concepts about measurements used as a basis for the set of practices described in section 3. Section 4 presents our conclusions.

2. Background

2.1. Basic Concepts

Some basic concepts are important to clarify why working on the definition of measures. First of all, measurement can be defined as follows: (i) The process by

which numbers or symbols are assigned to attributes of entities in the real world in such a way as to describe them according to clearly defined rules (Fenton and Pfleeger, 1997); and (ii) a set of operations having the object of determining the value of a measure ISO/IEC 15939 (2007).

In the first definition, an entity is an object (such as a person, a model, or a room) or an event (such as the testing phase of a software project). ISO/IEC 15939 (2007) summarizes that an entity is an object (a process, product, project, or resource) that is characterized by measuring its attributes. An attribute is a property of an entity (such as the color of a room, or the elapsed time of the test phase). The attributes are often defined using numbers and symbols (such as, a number of hours or the different labels for colors). Thus, we measure the attributes of entities by using specific measurement methods. The entity to be measured is the start point for a measurement definition. ISO/IEC 15939 (2007) organizes all these concepts associated with the measurement process defining some important concepts. It defends that a measurement process is driven by information needs (also named in literature as measurement goals), which are "insights necessary to manage objectives, goals risks and problems". To address the information needs, one can define:

- Base measures (named quality measure element in SQuaRE (ISO/IEC 25000, 2014)), that are independent measures defined in terms of an attribute (of an entity) and the method for quantifying it.
- Derived measures (named quality measure in SQuaRE (ISO/IEC 25000)), 2014) that are measures defined as a function of two or more values of base measures.
- Indicators - a measure defined using derived and base measures, and that is the basis for analysis and decision-making based on a model that combine one or more measures with associated decision criteria (thresholds or targets used to determine the need for action or further investigation, or to describe the level of confidence in a given result).

The methods used in the measures can be of two types: subjective, when the quantification of an attribute involves human judgment; or, objective, when the quantification is based on numerical rules such as counting, performed manually, or with automated tools. Finally, one of the following scales is associated with the measure (ISO/IEC 9126, 2001): nominal, ordinal, interval, or ratio. Measures using nominal or ordinal scales produce qualitative data, and measures using interval and ratio scales produce quantitative data (ISO/IEC 25000, 2014).

The literature is rich in measures for software products (models, code, software design, etc.), process and project. We can also find several systematic literature reviews (e.g. Nuñez-Varela et al. (2017), Carvalho et al. (2017), Hall et al., (2011), Bellini et al. (2008), Gómez et al (2008)) summarizing measures proposed in the last years.

2.2. Software Measurement Approaches

Several approaches have been proposed for the definition of measures. Some of the best known are: Goal-Question-Metric (GQM) (Basili *et al.*, 1994; Solingen and Berghout, 1999); the Goal-Question-Indicator-Metric (Park *et al.*, 1996), Practical

Software Measurement (PSM)[1] (McGarry *et al.*, 2002), the ISO/IEC 15939 (2007)) and the GQM/Metric Definition Approach (GQM/MDEA) (Briand *et al.*, 2002). All these approaches propose a set of steps in order to define software measurement for a product, process or project (see Figure 1). These approaches are differentiated by the number of steps presented and the detail given to carry out these steps. Some of them include some steps for the planning and integration of the measurement definition activities in the enterprise (see the first step of GQM, ISO/IEC 15939 and PSM in Figure 1). A common aspect in all these approaches is that they are goal-oriented, as introduced by GQM. Indeed, in a systematic literature review about software measurement, Tahir *et al.* (2016) concluded that the majorities of measurement planning models (83%) and measurements tools (90%) are extensions or improvement of GQM thanks to the goal-oriented measurement focus.

All these approaches are useful while defining measures and can be chosen without distinction, even if some (GQM, PSM and ISO/IEC 15939) are more used for measurement of software process and other (GQM, GQIM and GQM/MDEA) for software product (code, models, documentation etc.). By analyzing the steps particularly related to the software measurement definition (not considering the organization steps and the data collection), we note that three main issues are addressed in these approaches (Figure 3):

i. the **measurement goal definition** (steps highlighted in green - ♣ symbol in Figure 2), that focus on explicitly state the need for the measurement by formalizing the goal of the measurement in a clear and structured way;

ii. the **measure definition** itself (steps highlighted in blue - ♦ symbol in Figure 2), where, entities are identified and attributes are formalized via generic properties that characterize their measure (see previous section); and,

iii. the **measure evaluation** (steps highlighted in red -♠ symbol in Figure 2), that shows how to validate the measures defined and how to apply them for refinement and improvement of their definitions.

From our experience, regardless of which of the approaches from Figure 3 we follow, these issues are the most time-consuming and about which we have to overcome the main difficulties in the measurement definition (Dupuy-Chessa et al., 2014; Oliveira et al., 2012). The issue (i) should be deeply investigated by the team responsible for the measurements since it will guide the whole measurement program, being clear described, what is not always easy if we do not follow some standards template. Regarding measurement definition (issue (ii)), several difficulties are recognized: the need of an expert in the definition and interpretation of the measure, the insufficiency of adequate assessment tools, the definition and use of threshold and the large number of measures in literature. Finally, for the measurement evaluation (issue (iii)), validation procedures are not usual and experimental protocols are generally complex and difficult to define. However, we have to work with these difficulties and try to define as better as possible measurements that can support the quality evaluation of the software systems.

[1] New version of PSM v4.0b1 is available at http://www.psmsc.com/PSMGuide.asp

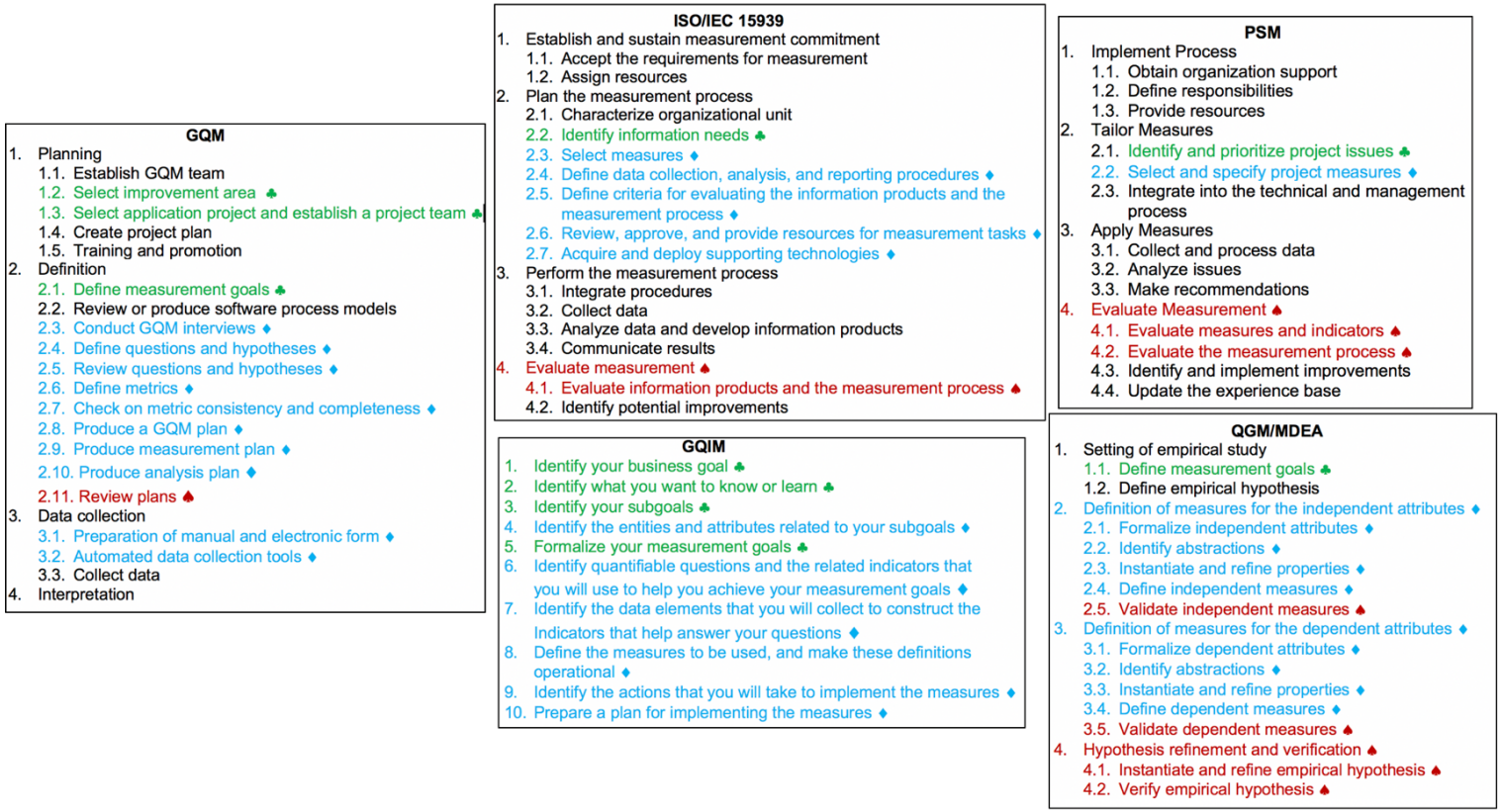


Figure 1. Steps of different software measurement approaches

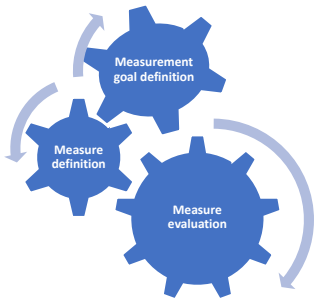


Figure 2. Main issues of software measurement approaches

3. Practices for Software Measurement Definition

Having been worked with measure definition for different proposes (e.g. ubiquitous systems (Carvalho *et al.*, 2018), legacy system (Ramos *et al.* 2004), interactive systems (Assila *et al.*, 2016; Gabillon *et al.*, 2013), web systems (Lima *et al.*, 2009) and software process improvement (Monteiro and Oliveira, 2010)), we have applied different practices that helped us to deal with the difficulties previously mentioned. These practices are presented in this section according to the main issues identified for software measurement presented in Figure 2.

3.1. Practices for Measurement Goal Definition

All approaches for software measurement definition starts by clearly establishing the measurement goal. In fact, measurement is a timing consuming and costly task. Therefore, the real need for the measurement should be identified since the beginning and should be revisited throughout the measurement definition to keep aligned with that goal. To that end, we should take into account the corporate objectives (Briand *et al.*, 2007), identify the business objectives that guide the organization's efforts and identify what one would like to know, in order to understand, evaluate, predict and improve activities related to the achievement of its objectives (Park *et al.*, 1996). A good practice is brainstorming with the team interested in the measurement trying to answer open questions, such as: What are the strategic goals of the organization/ project?, What are the major concerns (problems)?, What do we want to know/learn/improve?, How can we reach improvement goals?. After that, the measurement goal should be clearly written. We used as practice defining the goal following the structure defined in (Basili *et al.* 1994, Solingen and Berghout, 1999):

Analyze the *object* under measurement
For the purpose of understanding, evaluation, prediction, controlling, improvement
With respect to the *quality focus* of the object that the measurement focuses on
From the viewpoint of the *people* that measure the object
In the context of the environment in which measurement takes place.

Although it is a simple structure, each one of the lines of this definition is really important. From the beginning, it is essential to clearly define what is the object (the entity to be measured). As defined in section 2.1, the entity is the basis for the measurement process since we will measure the attributes of this entity.

For the *purpose* of the measurement, we should keep on mind what we want. It is very common to measure to obtain an accurate *characterization* or *evaluation* of the object for further analysis. *Prediction*, *controlling*, and *improvement* may require specific analysis models, comparisons among measurements, or continua data collection for some time slot.

Another important decision at this moment is the quality focus for the measurement definition. It is usually recommended to be decided according to the main interest of the organization by interviewing the sponsors. When in a research project, the choice regards the main interest of the project (e.g., usability, performance, etc.). In both cases, we usually use as practice look for a list of pertinent quality characteristic to support this decision. To that end, ISO/IEC 25010 (2011) can be used since it presents a set of quality characteristics (13) divided into sub-characteristics (42) related to outcomes of interaction with a system (the quality in use model) and related to the system/software product quality properties (the product quality model). Since this standard can be applied to any kind of software system, it is a rich source of quality focus when the entity to be evaluated is a product. However, a good practice is also look in the literature searching quality characteristics specific for the object been measured. This is particularly important when dealing with new kind of software systems like ubiquitous applications, IoT applications (e.g., smart cities), conversational agents, etc. Systematic mappings studies (Petersen *et al.*, 2015) proven to be quite useful in this case. It may reveal different quality focus not presented in the standards, and that covers particularities of that kind of system. Our study for ubiquitous systems (Carvalho *et al.*, 2018) showed particular quality focus regarding context-awareness, transparency, calmness, attention, and mobility.

Finally, it is important to define the *point of view* we are interested in precisely. That means the people that will measure and will benefit from the measurements. This will guide the definition of the measures to decide which attributes of the entity we are interested in measure to the purpose defined.

Figure 3 presents different examples of measurement goal definitions. It is worth mentioning some peculiarities of each of these definitions, as follows:

- in (a), the business goal that motivates the measurement was to evaluate an existing legacy system, in a short time in order to support outsourcing companies to establish the maintenance contracts to other companies. We aimed to have a view about the documentation and code of the legacy system to have an idea of the amount of work required.
- In (b), the interest was collecting data to define indicators for a service catalog of an SLA (Service Level Agreement) about website accessibility in order to support the definition of SLAs by the Brazilian Government while contracting companies to develop and maintain their web sites;

- In (c), the motivation was to evaluate quality characteristics that impact directly on the human-computer interaction in mobile applications;
- In (d), the context of transport applications imposes that the user interface should be as simple and complete as possible to support decision-making in real-time;
- In (e), the business goal behind this measure is to control the software development to monitor the time and cost to be aligned with the established contracts for a service in a software house.

Analyze the system documentation **for the purpose** of assessing **with respect to** completeness and consistency **from the viewpoints of** analysts and programmers **in the context of** the outsourced maintainer.	**Analyze** the system source code **for the purpose of** assessing **with respect to** the complexity to understand and modify it **from the viewpoints of** analysts and programmers **in the context of** the outsourced maintainer.

(a) Measuring legacy systems (Ramos *et al.*, 2004)

Analyze content on websites **for the purpose of** evaluation **with respect to** accessibility, **from the viewpoint of** the user with visual disabilities **in the context of** governmental services	**Analyze** ubiquitous systems **for the purpose of** evaluating **with respect to** context-awareness, mobility, transparency, attention and calmness **from the viewpoint of** user and developer **in the context of** mobile applications

(b) Measuring web applications (Lima *et al.*, 2009)

(c) Measuring ubiquitous applications (Carvalho *et al.*, 2018)

Analyze user interface **for the purpose of** improving **with respect to** information density **from the viewpoint of** developer **in the context of** transport application	**Analyze** project scope (time and cost) **for the purpose of** controlling **with respect to** performance **from the viewpoint of** managers **in the context of** software houses**.**

(d) Measuring interactive systems (adapted from (Assila *et al.*2016))

(e) Measuring software process (adapted from (Monteiro and Oliveira 2011))

Figure 3. Examples of measurement goal definition

3.2. Practices for Measure Definition

The measurement definition is the core of any measurement approach. First of all, it is important to be aware that there are a lot of measures already defined and being used in the literature. Consequently, a practice that we have applied is to look for measures in the literature considering the quality focus we have defined in the measurement goal before starting any activity for defining measures. Again, systematic mapping study is quite adequate. However, although we can find measures in literature, they are not always described and formalized in a way that we can reuse it. For instance, in the systematic mapping for ubiquitous application we have done (Carvalho *et al.*, 2018), we found 218 measures, but analyzing them against the formalization of measures purposed by SQuaRE (ISO/IEC 25000, 2014) and

(ISO/IEC 15939, 2007) we note that only a small part (39) was well defined presenting measurement functions and quality measurement elements for a definition. Nevertheless, even in the case we do not have the complete description of the measure, to have a list or ideas of measures for the quality focus at hand is very helpful.

By looking the literature, we can also find several measures to measure the same quality characteristics (for instance, we found several measures for code size and complexity of the code while evaluating legacy systems (measurement goal (a) in Figure 4). Therefore, we can either choosing one to work with or applying all of them to have different perspectives on the quality focus being measured.

To define the new measures, two main points should be considered. The first one is to have the entity to be measured as the main element. That means we should consider the entity, explore and define all attributes this entity has, and consider other entities (and also their attributes) that have some impact/relation on the entity in the study. All this information will be useful while defining the measure. The second point is to consider experts in the kind of object being measured. It is worth to choose different experts and perform individual and group interviews with them. Since group interviews are not easy to organize (problem of schedule and availability), it is good to have some support to integrate opinions and guide the measure definition. To that end, a good practice we have applied is to use abstraction sheets (Solingen and Berghout, 1999). This document summarizes the key issues about the measurement goal into four parts as follows:

- **Quality Focus** - in which experts should choose or define possible measures for the defined quality focus. At this moment, the list of measures we collected from the literature (even with only names of the measure and not a complete formalization) is really useful. They work like insights for the experts, promote discussion and stimulate the definition of measures;
- **Baseline Hypotheses** - in which experts use their experiences (from other projects) to set a possible interpretation value that means what they expect to find as acceptable values. We consider that at least one baseline hypotheses for each measure should be defined;
- **Variation Factors**, in which experts identify potential factors that can impact the suggested possible measures; and,
- **Impact of Variation Factor**, in which experts should answer how the various factors could impact the measures and what kind of dependencies exists between the measures and the factors.

Figure 5 presents an abstraction sheet regarding the measurement goal (c) presented in Figure 4.

With the abstract sheet filled in, we can (with the experts, if possible) define the measures starting from the measures listed in the quality focus quadrant, and then trying to define a measure for each variation factor that impact those measures. The definition of those measures should be based on the attributes of the entities related to them. From or experience, abstract sheets are effective in defining measures, even if we have only one expert or a study group of researchers interested on the entity being evaluated since it helps to brainstorm and to structure what can be measured.

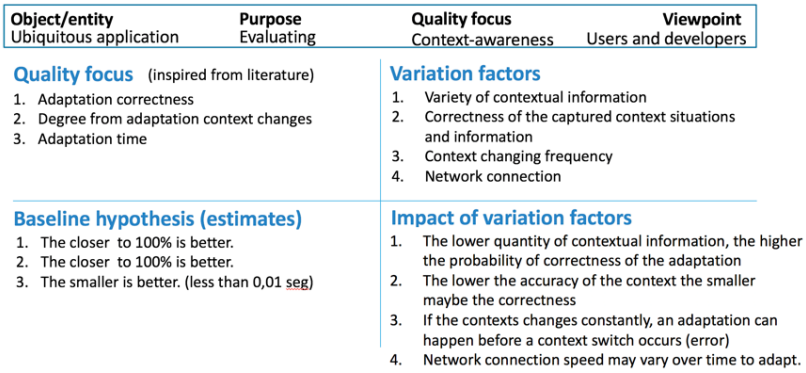


Figure 4. Example of abstraction sheet for ubiquitous applications evaluation (adapted from (Carvalho et al., 2018))

When defining a measure, we can conclude that to better evaluate the quality focus in study is necessary the combination of two measures in the same view. To that end, indicators as defined by ISO/IEC 15939 (2007) can be applied. For instance, for evaluating the information density of interactive systems, we concluded that it was important to have measures not only about the user interface itself but also about the users' opinion concerning their perception of the density of information while executing some tasks (Assila *et al.*, 2016). Indicators can also give some prevision for the object been evaluated. For instance, to control the software project (measurement goal (e) in Figure 4), we crossed measures of cost and time in control charts that support the analysis of stability over time.

To support the measure definition, a good practice is to follow a template for the measurement description. ISO/IEC 25000 series and ISO/IEC 15939 (2007) provide a list of elements we should consider while describing the measures. In general, at least the following information must be provided:

- Name – defined for convenience and that express the main meaning of the measurement;
- Description of the measure – the information described by the measure or gathered for the measures. This description must be cleared enough to make an easy understanding of the measured goal. To that end, we can use a sentence and/or a question to be answered by the application of the measure;
- Measurement function – ISO/IEC 25022 (2012) defines as an equation showing how the quality measure elements (base measures) are combined to produce the quality measure (derived measure). ISO/IEC15939 (2007) states that it can also be an algorithm or calculation performed to combine two or more base measures.
- Interpretation – a description to support the interpretation of the result for decision making. We propose two general practices. The first one is to normalize the value of the measure within 0.0 to 1.0 and that consider the interpretation as the closer to 1.0 is better (as suggested by in ISO/IEC 25022 (2012)) or close to

0.0 is better. The second practice is to apply the baselines defined in the abstraction sheets to define the initial values for thresholds. Those values can be revised after empirical evaluations (see next section).

- evaluation method – "procedure describing actions to be performed by the evaluator in order to obtain results for the specified measurement applied to the specified product components or on the product as a whole" (ISO/IEC 25000, 2007). In practice, we used three kinds of evaluation methods, as follows:
  - Questionnaires – classical way to collect data for the subjective measure from users' opinion. In this case, the Likert scale is commonly used. We have also used a continuous scale named VAS (Visual Analogue Scale) that allows the application of a wider range of statistical methods to the measurements.
  - Third-party observation – a third person that observe users interacting with the system during an evaluation session. The third person takes notes for further analysis. Forms with the data to be collected during the observation should be provided. The evaluation sessions can be face-to-face or remote.
  - Interaction log – use of automated data collection tools (usually specifically implemented for the system to be evaluated). It collects a trace of the execution of the system with specific data previously defined. The data to be collected is defined based on the defined measures.

Moreover, specific functions can be coded to collect measures from code (e.g.: complexity of the code). Several plug-ins and open-source tools are also available to this purpose[2].

Table 1 shows two examples of measure description. The first one was defined based on the abstraction sheet presented in Figure 4. Several examples of indicators description following ISO/IEC 15939 (2007) structure can be found in (Assila *et al.*, 2016; Monteiro and Oliveira, 2011).

3.3. Practices for Measure Evaluation

In order to guarantee the validity of a defined measure, we should assure its theoretical correction and apply it in several applications. The ideal scenario would be to have a historical database of the measures collected and use it to refine the thresholds defined to support the interpretation. This scenario requires a long-term research. In any case, a common sense in literature is that two kinds of validation are necessary for measurements (Srinivasan and Devi, 2014): theoretical and empirical.

Theoretical validation aims to confirm that the measurement does not violate any necessary properties of the elements of measurement (Srinivasan and Devi, 2014). To that end, a theory of definition of measures must be applied. We have used the theory of measures proposed in (Kitchenham *et al.*, 1995), (Fenton and Pfleeger, 1997) and

[2] For instance, for code in java (https://www.spinellis.gr/ sw/ckjm/; https://github.com/mauricioaniche/ck; https://www.sourcemeter.com); in C/C++, C# and Python (https://www.sourcemeter.com)).

(ISO/IEC 25000, 2007). In general, the main theoretical items from all these theories are the following: the entity to be measured, the attribute (also called "property to quantify" in (ISO/IEC 25000 series)), the scale type, and a measurement unit. Kitchenham *et al.* (1995) defend that it is also important to consider the adequation of the instrument and the adoption of a measurement protocol. (Srinivasan and Devi, 2014) also shows that mathematical properties may be assured. In any way, the main idea is to assure that we followed a theory of measurement while defining the measures. Table 2 shows a simple way of presenting the theoretical validation, by explicitly defining all elements of the measures (examples used in this paper). In the same way, Cheikhi *et al.* (2014) presents a theoretical validation of traditional well-known measure for object-orient systems.

Table 1. Example of specification of measurements

Name	Description	Measurement function	Interpretation	Evaluation Method
Adaptation Correctness (Carvalho *et al.*, 2018)	Does the adaptation occur correctly in the current context of the user?	$X = \frac{\left(\sum_{j=1}^{N} \frac{Aj}{Bj}\right) * 100}{N}$ N=Number of different adaptations Aj =Number of correctly performed adaptations j Bj =Number of performed adaptations j	The closer to 100%, the better.	Interaction log and Third party observation
Overall density (for graphical user interface) (Adapted from (Assila *et al.*, 2016))	It measures the percentage of display used to present all information.	X= Used space / Total space of an interface	0.0 < X < 1.0 The closer to 0 the better. **or** Acceptable: for values ϵ]0%,X] Unacceptable: for values ϵ]X;1] X in literature is usually 0.25-0.3	Coded function[3]

The empirical validation aims to confirm that a measure has the desired predictive power for predicting or evaluating the variable of interest (Antinyan *et al.*, 2016). According to (Srinivasan and Devi, 2014), three types of empirical validations are: surveys, case studies and experiments. Preferably, case studies, and experiments should be performed in a variety of application domains.

[3] This function in general calculate the total area of all the graphical components displayed in the interface against the overall area of the interface.

Table 2. Example of theoretical elements of a measurement

Measure	Base measure	Entity	Attribute	Scale	Unit
Adap-tation Correct-ness	Number of different adaptations	Running system	Adaptation Types (*e.g.*, adaptation occurred by battery, adaptation occurred by location)	Ratio	Adap-tation
	Number of correctly performed adaptations	Running system	Correctly performed adaptations	Ratio	Adap-tation
	Number of performed adaptations	Running system	Performed adaptations	Ratio	Adap-tation
Overall density	Used space of an interface	GUI compone nts (labels, textfields, images, etc.)	components' height components' width	Ratio	Pixel
	Total space of an interface	GUI	height width	Ration	Pixel

We have been using the action research method proposed by Antinyan *et al.* (2016) (Figure 5). In general, the measures are selected, calculated, evaluated, and redefined based on the evaluation until they are perceived to be good measures. To that end, the designer of the measurement (who defined the measures) and a reference group (a group of practitioners who work closely with the artifacts that are to be measured) work together during the cycle of validation.

The empirical evaluation should start with the definition of the research protocol (for a case study or experiment) to be applied. Classical elements such as subjects, time and environment for the study, and data collection methods should be clearly described. After the execution of the case study/experiment and the data collection, the designer presents the results for a reference group, and they brainstorm together to understand how effectively the selected measure can assess the variable of interest (Antinyan *et al.*, 2016). The aim is to evaluate if a given measure is effective or not and to check with the reference group whether the measurement results match the current state of the application (that means, the results correspond to what they know about the entity been evaluated). As consequence, either the reference group agrees with the results achieved or disagrees, indicating possible changes to be made. All agreements, disagreements, and reasons should be registered in a document that will be used for the measurement improvement. Then, a new cycle of evaluation is performed. From our experience, typical suggestions from the reference group are the

following: the modification of thresholds, modification of interpretation procedures and improvement of collect procedures (for instance, redefinition of questionnaires, and inclusion of new facilities in the automated tools).

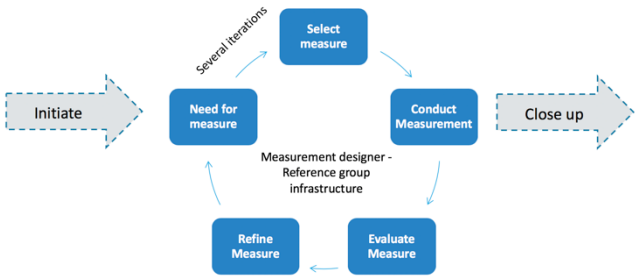


Figure 5. Action research cycle for validation of measures (Antinyan et al., 2016).

From our experience, we observed that after some action research cycles, we can obtain more stable thresholds for future interpretations. To deal with the complexity of definition of experimental protocols, we suggest the application of short cycles of evaluation with simple protocols, that can be evolved after each cycle.

4. Conclusion

This paper presented a set of practices to be applied while defining measures. These practices can be applied with any measurement approach while defining the measurement goal and measures, and proceeding the measure evaluation.

We can summarize the defined practices drawing the following principles: (i) follow a measurement approach; (ii) use a clear template for the measurement goal definition; (iii) consider the entity to be measured as the base element for measurement definition (we recall that we measure attributes of entities); (iv) look for measures already defined in literature even if not formalize is better than start from scratch; (v) include experts (developers of the kind of applications, managers, technical team, etc.) in the definition and the validation of the measures; (vi) formalize measurements by defining the measure name, description, measurement function, method, unit of measurement, etc.; (vii) automatize as much as possible (to collect the data for the base measures); (viii) apply and validate measurements.

We consider that the practices presented in this work can be directly applied to new projects and we hope that they can motivate the measurement definition and dissemination for new kinds of software system.

Acknowledgements

We strongly thank all co-authors of papers listed in this article.

References

Antinyan V., Staron M., Sandberg A. (2016). Validating Software Measures Using Action Research - A Method and Industrial Experiences, *17th International Conference on Enterprise Information Systems*, vol. 2, p. 15–27.

Assila A., Oliveira K., Ezzedine H. (2016). Integration of Subjective and Objective Usability Evaluation based on ISO/IEC 15939: a Case Study for Traffic Supervision Systems. *International Journal of Human-Computer Interaction*, 32 (12), p. 931-955.

Basili, V., Rombach, H. (1994). Goal Question Metric Paradigm, Encyclopedia of Software Engineering. Encyclopedia of Software Engineering – 2.

Bellini C.G., Pereira R.D.C.D.F., Becker J.L. (2008). Measurement in software engineering from the roadmap to the crossroads. *International Journal of Software Engineering and Knowledge*, 18(1), p. 37–64.

Briand L.C, Morasca S., Basili V. (2002). An Operational Process for Goal-Driven Definition of Measures. IEEE Transactions on Software Engineering, vol. 28, no. 12, p. 1106-1125.

Carvalho R., Andrade R., Oliveira K. (2018). AQUArIUM - A Suite of Software Measures for HCI Quality Evaluation of Ubiquitous Mobile Applications. *Journal of Systems and Software*, vol. 136, p. 101-136.

Carvalho R., Andrade R., Oliveira K., Santos I., Bezerra C. (2017). Quality characteristics and measures for human-computer interaction evaluation in ubiquitous systems. *Software Quality Journal*, 25(3), p. 743-795

Cheikhi, L., Al-Qutaish, R.E., Idri. A., Sellami, A. (2014) Chidamber and Kemerer Object-Oriented Measures: Analysis of their Design from the Metrology Perspective, International Journal of Software Engineering and Its Applications, vol.8, no.2, p. 359-374.

Dupuy-Chessa S., Oliveira K., Si-Said cherfi S. (2014). Qualité de Modèles : retour d'expérience. XXXIIème INFORSID, p. 363-378.

Evers V., Cramer H., Van Someren M., Wielinga, B. (2010). Interacting with adaptive systems. *Interactive collaborative information systems*, p. 299–325.

Fenton, N., Pfleeger, S. *Software Metrics A Rigorous & Practical Approach*, 2nd. Ed., PWS Publishing Company, 1997.

Gabillon Y., Lepreux S., Oliveira K. (2013) Towards ergonomic User Interface composition: a study about information density criterion. *15th International Conference on Human-Computer Interaction*, p. 211-220.

Gómez O., Oktaba H., Piattini M., García, F. (2008). A systematic review measurement in software engineering: state-of-the-art in measures. ICSOFT, LNCS 5007, p. 165–176.

Hall T., Beecham S., Bowes D., Gray D., Counsell S. (2011) A Systematic Literature Review on Fault Prediction Performance in Software Engineering. *IEEE Transactions on Software Engineering*, 38(6), p. 1276-1304.

ISO/IEC 15939. *System and Software Engineering – Measurement Process*, 2nd edition, 2007.

ISO/IEC 25000. (2014) *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE)*.

ISO/IEC 25010. (2011) -*SQuaRE – System and Software Quality Models*.

ISO/IEC 25022. *SQuaRE) — Measurement of quality in use*, July 2012.

ISO/IEC 9126. *Software engineering - Product quality - Part 1: Quality model*, 2001.

Jafari S., Mtenzi F., O'Driscoll C., Fitzpatrick R., O'Shea, B. (2011). Measuring privacy in ubiquitous computing applications. *International Journal of Digital Society*, 2(3), p. 547–550.

Kitchenham, B., Pfleeger, S.L., Fenton, N., 1995. Towards a framework for software measurement validation, *IEEE Transactions on Software Engineering*, pp. 929–944.

Lee, J., & Yun, M. H. (2012). Usability assessment for ubiquitous services: Quantification of the interactivity in inter-personal services. *IEEE international conference on management of innovation & technology*.

Lima S., Lima,F., Oliveira K. M. (2009) Evaluating the Accessibility of Websites to Define Indicators in Service Level Agreements, *11th International Conference on Enterprise Information Systems*, p. 858-869.

McGarry J, Card D, Jones C, Layman B, Clark E, Dean J, Hall F. (2002) *Practical Software Measurement: objective information for decision makers*. 1st ed. Addison-Wesley: Boston.

Monteiro L., Oliveira K. (2010). Defining a catalog of indicators to support process performance analysis. *Journal of Software Maintenance and Evolution: Research and Practice*, 23 (6), p. 395-422.

Nuñez-Varela, A.S., Pérez-Gonzalez, H.G., Martínez-Perez, F.E., Soubervielle-Montalvo C. (2017) Source code metrics: A systematic mapping study, *Information and Software Technology*, vol. 128, p. 164-197.

Oliveira K., Thion V., Dupuy-Chessa S., Gervais M.-P., Si-Said cherfi S., Kolski C. (2012). Limites de l'évaluation d'un système d'information : une analyse fondée sur l'expérience pratique. Actes XXXème Congrès INFORSID, p. 411-427.

Park R.E., Goethert W.B. e Florac W.A. (1996). *Goal Driven Software Measurement – a Guidebook*, CMU/SEI-96-BH-002, Software Engineering Institute.

Paschou M., Sakkopoulos E., Sourla E., Tsakalidis, A.(2013). Health Internet of Things: Metrics and methods for efficient data transfer. *Information and Software Technology*, 34, p. 189-199.

Petersen, K., Vakkalanka, S., & Kuzniarz, L. (2015). *Guidelines for conducting systematic mapping studies in software engineering: An update. Information and Software Technology*, vol. 64, p. 1-18.

Ramos C. S., Oliveira K. M., Anquetil, N. (2004). Legacy Software Evaluation Model for Outsourced Maintainer. *8th IEEE European Conference on Software Maintenance and Reengineering*, p. 48-57.

Solingen, R. van, Berghout, E. (1999). *The Goal/Question/Metric Method: A practical guide for quality improvement of software development*. McGraw-Hill.

Srinivasan, K.P., Devi, T. (2014). Software Metrics Validation Methodologies in Software Engineering. *Journal of Software Engineering and Applications*, vol. 5, p. 87–102.

Tahir T., Rasoola G., Gencelb C., (2016). A systematic literature review on software measurement programs. *Information and Software Technology*, 73, p. 101–121.

Wu, C. L., & Fu, L. C. (2012). Design and realization of a framework for human–system interaction in smart homes. *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, 42(1), p. 15–31.

A Unified Vision of Configurable Software

Houssem CHEMINGUI [1,2], Inès GAM [1,2], Raúl MAZO [3, 4], Henda BEN GHEZALA [2], Camille SALINESI [1]

1. CRI Laboratory, Paris 1 Panthéon Sorbonne University, 90 rue de Tolbiac, 75013 Paris - France
2. RIADI Laboratory, ENSI Manouba University, 2010 Manouba - Tunisia
3. Lab-STICC, ENSTA Bretagne, 2 rue François Verny, Brest - France
4. GIDITIC, Universidad Eafit, Carrera 49 N° 7 Sur-50, Medellin - Colombia

RÉSUMÉ. En pratique, la configuration logicielle est une tâche difficile et sujette à erreurs en raison du grand nombre d'exigences et de contraintes à satisfaire simultanément. Les parties prenantes se trouvent confrontés à des problèmes de rigueur et de passage à l'échelle lors de la configuration de logiciels: les méthodes employées pour spécifier les systèmes à configurer ne permettent pas de maîtriser de manière formelle et systématique un nombre important de décisions complexes. Cependant, des approches visant à surmonter ces verrous existent et ont été publiées, mais dans certains domaines; peuvent-elles être adaptées pour toute sorte de logiciels configurables? Les défis scientifiques de la configuration logicielle peuvent ils ainsi être transposés? Cet article aborde ces questions à travers un cadre unificateur de configuration identique et adaptée à différents cas d'utilisation et contextes.

ABSTRACT. In practice, software configuration is error-prone due to the plethora of requirements and constraints to satisfy at the same time. Practitioners face awkward scalability issues when configuring large variability-based software. Indeed, standard variability modeling methods such as feature and even decision models fail in mastering a suitable configuration process implying a huge panel of complex decisions. However, published solutions, aiming to overcome these obstacles, exist but they have been designed in separate ways; can they be adapted for all sorts of configurable software? Can the scientific challenges of software configuration be transposed? This paper addresses these issues through a unified framework of configuration encompassing different use cases and contexts.

Mots-clés : Configuration logicielle, passage à l'échelle, défis scientifiques, cadre unificateur

KEYWORDS: Software configuration, scalability issues, scientific challenges, unified framework

DOI:10.3166/RCMA.25.1-n © 2020 Lavoisier AR_DOI

1. Introduction

Developing configurable software such as ERP, SPL or COTS aims to boost productivity and thus maximize revenues. However, uncareful management of the configuration process (Mittal et al., 1989) creates serious vulnerability and

inconsistency problems in the software. The huge number of decisions and incomprehensible interactions between them present a major problem. Unfortunately, this problem is common for all configurable software systems even worse in large scale variability based software (Chemingui et al., 2019).

Currently, a panoply of research solutions aiming to improve the configuration process exists in literature, nevertheless, these solutions are designed in separate ways. For instance, solutions that master SPL configuration fail to be adapted to master COTS configuration. Therefore, thinking about a systematically extending and adaptation of solutions to other kinds of configurable systems of software leads this research to the following research questions:

RQ1: What are the open research challenges of unified software configuration?

RQ2: What are the common configuration assets of configurable systems?

This paper advocates a unified vision that can be formalized through a conceptual framework considering all sorts of configurable software in a consistent way. The motivation of this unified vision is to address common configuration issues and challenges with an identical series of methods, techniques and tools. The paper recognizes that in practice software configuration occurs in specific contexts, with different problems, goals, applications and use cases.

2. Challenges of Software Configuration

Considering all sorts of configurable software conjointly, leads this research to be focusing on topical and thought-provoking challenges related to design solutions and scientific research potentials.

2.1. Design solutions and industry need

In practice, configuration process models fail to scale up to the huge panel of variants that are encountered in real life software configuration settings. For instance, it becomes impractical to make an extensive list of all possible decision processes leading to final configurations of an ERP. Even worse, it is quickly impractical to match these decisions with the actual collection of requirements of the future users. At the level of an enterprise, collections can be large and up to several thousands. In fact, major issues of configurable software are related to scalability enforcement and control. Besides, stakeholders usually misunderstand the configuration variants, their semantics and their independencies. This unfortunately common issue in configurable software can be formalized as the mismatch. In the case of COTS, this mismatch in form, topic, content, and level of abstraction between the requirements causes complex matching problems (Zoukar et al., 2004). Failing to properly match users' and COTS requirements results is failing to meet organizational expectations, to user dissatisfaction, and in the end to high risks of

project failure. In practice, COTS implementation guides hardly provide any support to guide the design of a suitable solution that takes this issue into account in an equilibrated way. The solutions that are proposed to guide the configuration process are most often driven by the solution; other approaches such as decision models, and goal driven approaches are still subject of research. A compendium about scalability issues of configurable systems are discussed in our previous works (Chemingui et al., 2019). In other works, we showed that it is also possible to address scalability issues in a similar way to the approaches used for SPL (Software Product Lines) configuration (Mazo et al., 2014).

In this respect, configurable systems need a deeper focus on their design and user understanding, not solely to know who they are, but to dive deeper into their motivations and their exhibited behaviors. A matching between user requirements, context information and the system model seems mandatory. As far as we know, there are no efficient software resources that blend this matching with the different configuration areas. It seems interesting to think about a software standard with a universal dimension that is able to overcome scalability and flexibility problems. Independently of its nature, this solution can be a programming language, a design notation or a complete method that is able to support broaden system contexts and particular user properties.

According to our previous experiences with industrials in automotive (Dumitrescu et al., 2013), electronics (Triki et al., 2015) and health (Djebbi et al., 2007) domains, we suggest that expected guidance solutions should consider design fuzziness to meet the user satisfaction policies. For example, what principles should be taken into account throughout the design? How configuration information and constraints can be better presented? What users need to a better understanding of complex configuration alternatives? Regardless of the size of variants, design has to support easy but not abstract data allowing to infer rapidly the plethora of configurations when high constraints occur. On the one hand, a focus on the delivery time will have a significant impact in the configuration management. Moreover, design has to provide simplified mechanisms to reveal configuration use cases and scenarios. It is highly recommended to follow domain processes, divide tasks and create views automatically from a big range of artefacts and interests. On the other hand, the solution supporting tool will provide simple ways to express variants and their interdependencies allowing an easier process of configuration. For instance, incorporating Human Machine Interface (HMI) solutions such as dynamic forms and 3D views increases the user comfort and decreases the configuration complexity.

2.2. Scientific Research Potentials

From an organizational point of view, a collaborative research focus may conduct the investigation of cited challenges in real environments. Methods such as Participatory action research (Chevalier et al., 2019) and ethnography Research (Holland et al., 2004) are recommended to achieve a better understanding of

configuration design and potential users interests. Too often, research involvement allows to motivate real technology underpinnings and its usefulness to overcome scalability problems in this case. In fact, there is a significant gap between what industry needs and what academic research focuses on. Usually, the interaction between industrial actors and academics faces several barriers since companies keep secret their strategic activities. Consequently, the fear of breaching confidentiality impacts significantly the research valorization. The lack of collaboration may cause trustworthiness shortcomings making it more complex to valorize research outcomes and relate best practices. As a result, proposed methods and tools coming from academic research are making the same mistakes with unforeseen incompatibilities.

3. Common considerations of Software Configuration

At first sight, suitable solutions for all configurable software need to (i) provide exact information about product constituents and their interactions; (ii) meet non-functional properties such as performance, ease of use and robustness, (iii) forcecast information even about next constituents to configure.

Our expertise with very large variability systems such as Electronic Parking Brake Systems, the Automatic Lighting System, or the French railways operated by SPLs, COTS and ERPs, showed that design methods are still suffering from scalability. The most widely used variability modeling methods were *feature models* (Kang et al., 1990) *use cases* (Jacobson et al., 1993), *decision models* (Quinlan, 1990), *goal models* (Barron et al., 2001) and *constraint programming* (Rossi et al., 2006). In front of scalability problems, resolution mechanisms including alternative structures of the model and estimation of the configuration errors were proposed to conduct a significant software reconfiguration (Giese et al., 2006). The configuration process can be streamlined by a series of interfaces aiming to a prealable preparation of a simple process, but a large number of variants still impossible to manage. Adding to that, software approaches such as feature toggles (Schermann et al., 2018) and versioning (Sigal et al., 1999) were widely adopted by developers to control combinatorial explosions. The main usage of these approaches is to avoid conflict that can arise when merging changes in software. Although this also can lead to technical debts that arise due to constraints violation after a feature has been switched on/off. Frequently, constraint violations disturb other parts of the system.

However, the literature reveals a prominent focus on configuration considerations but in different ways that are specific to each software kind, while in fact the deep problems are similar in nature. The idea put forward is to focus on the common configuration considerations, think about a unified conceptual framework propeling potential generic solutions. It seems interesting to emphasize, that independently of the variability modeling method and the kind of software you are dealing with, there are common considerations of configuration. The landscape to sketch aims to substantiate that software configuration handles the same assets independently of variability modeling methods. In the aforementioned variability methods,

composition, and order present recurrent activities' patterns during a configuration process. Moreover, one of the key principles of the configuration processes is to guarantee that all *domain constraints* are verified. In fact, variability constraints can have cascading effects during configuration processes, and they are the source of complex problems that raise exponentially with the number of optional assets in all variability based systems. For example, a SPL Eshop can be composed of several functionalities such as a search menu, a payment system and a security mode. Purchases can be paid through a credit card and/or a bank transfer. Security mode can be High or Standard. Selecting the credit card system requires a high security mode. In instance, configuring the security mode before selecting the payment system is error prone and leads to undo choices. Consequently, following a configuration order that respects the constraint domains is necessary. In fact, when the number of the eshop functionalities increases, the number of potential eshops increases and user decisions to be made increases also.

4. Conclusions

One of the challenging problems facing the software configuration community today is how to unambiguously define solutions for scalability in such a way to achieve consistency, flexibility and preference considerations. Furthermore, how to define solutions in a rich, generic and unified form allowing to consider all sorts of configurable software by combining advantages and avoiding drawbacks? All knowledge acquired in the different fields of configuration enhancement must be extended, adapted, and shared in a transparent manner. Therefore, there is a clear need to ease the cooperation of scientifics and industrials by creating a common ground of concepts and knowledge sharing. To meet these challenges, an efficient research valorization is needed to meaningfully conduct a collaborative enhancement of the configuration process **[RQ1]**. With the wisdom of hindsight, this vision paper assumes that all sorts of configurable software can be analyzed and improved in a unified way. Generally, configurable software aims to reach valid products implying common configuration assets to deal with namely composition, order and domain constraints and obviously other particular assets such as subtyping and cardinalities **[RQ2]**.

This research is under deep technical experiments that are supported by the PHC UTIQUE project Nº 16G/1416 (called CONFIGURE) and the european project Nº 15010 (called REVaMP²).

References

Astesana, J. M., Cosserat, L., & Fargier, H. (2010, October). Constraint-based vehicle configuration: A case study. In 2010 22nd IEEE International Conference on Tools with Artificial Intelligence (Vol. 1, pp. 68-75). IEEE.

Barron, K. E., & Harackiewicz, J. M. (2001). Achievement goals and optimal motivation: testing multiple goal models. Journal of personality and social psychology, 80(5), 706.

Chemingui, H., Gam, I., Mazo, R., Salinesi, C., & Ghezala, H. (2019). Product Line Configuration Meets Process Mining. In Proceeding of the 11th International Conference on ENTERprise Information Systems (pp. 199-210)

Chevalier, J. M., & Buckles, D. J. (2019). Participatory action research: Theory and methods for engaged inquiry. Routledge.

Djebbi, O., Salinesi, C., & Diaz, D. (2007) Product Line Requirements Matching and Deriving: the RED-PL Guidance Approach.

Dumitrescu, C., Mazo, R., Salinesi, C., & Dauron, A. (2013, August). Bridging the gap between product lines and systems engineering: an experience in variability management for automotive model based systems engineering. In Proceedings of the 17th International Software Product Line Conference (pp. 254-263). ACM.

Giese, H., & Tichy, M. (2006, September). Component-based hazard analysis: Optimal designs, product lines, and online-reconfiguration. In International Conference on Computer Safety, Reliability, and Security (pp. 156-169). Springer, Berlin, Heidelberg.

Holland, D., & Leander, K. (2004). Ethnographic studies of positioning and subjectivity: An introduction. Ethos, 32(2), 127-139.

Jacobson, I. (1993). Object-oriented software engineering: a use case driven approach. Pearson Education India.

Kang, K., Cohen, S., Hess, J., Novak, W., Peterson, S., 1990. Feature-Oriented Domain Analysis (FODA) Feasibility Study. Technical Report CMU/SEI-90-TR21, SEI, Carnegie Mellon University

Mazo, R., Assar, S., Salinesi, C., & Hassen, N. B. (2014). Using Software Product Line to improve ERP Engineering: literature review and analysis. Latin American Journal of Computing Faculty of Systems Engineering Escuela Politécnica Quito-Ecuador, 1(1), 10.

Mittal, S., & Frayman, F. (1989, August). Towards a Generic Model of Configuration Tasks. In IJCAI (Vol. 89, pp. 1395-1401).

Quinlan, J. R. (1990). Decision trees and decision-making. IEEE Transactions on Systems, Man, and Cybernetics, 20(2), 339-346.

Rossi, F., Van Beek, P., & Walsh, T.. (2006). Handbook of constraint programming. Elsevier.

Schermann, G., Cito, J., & Leitner, P. (2018). Continuous experimentation: challenges, implementation techniques, and current research. Ieee Software, 35(2), 26-31.

Sigal, A. D., Bien, D., & Pissarra, A. (1999). U.S. Patent No. 5,881,292. Washington, DC: U.S. Patent and Trademark Office.

Triki, R., Salinesi, C., & Mazo, R. (2015, May). Three strategies to specify multi-instantiation in product lines. In 2015 IEEE 9th International Conference on Research Challenges in Information Science (RCIS) (pp. 211-216). IEEE.

Zoukar, I., & Salinesi, C. (2004, April). Matching ERP Functionalities with the Logistic Requirements of French Railways: A Similarity Approach. In ICEIS (3) (pp. 444-450).

Modélisation graphique des environnements proxémiques basée sur un DSL

P. Pérez [1], P. Roose [1], M. Dalmau[1], Y. Cardinale [2], D. Masson[3], N. Couture[4] ,

1. Université de Pau et des Pays de l'Adour, LIUPPA, France
paulo.perez-daza,, Philippe.Roose, marc.dalmau @iutbayonne.univ-pau.fr
2. Universidad Simón Bolívar, Venezuela
ycardinale@usb.ve
3. DEV 1.0 , France
d.masson@dev1-0.com
4. Univ. Bordeaux, ESTIA INSTITTUTE OF TECHNOLOGY, LaBRI, France
n.couture@estia.fr

RESUME. L'interaction proxémique est un domaine émergent dont l'objectif est d'améliorer l'expérience utilisateur. La proxémique est basée sur cinq dimensions (Distance, Identité, Localisation, Mouvement et l'Orientation que nous appelons DILMO) pouvant être utilisées pour définir des interactions entre personnes et/ou appareils numériques. Les travaux de recherche et développement actuels dans ce domaine sont essentiellement axés sur le développement d'applications proxémiques via une boite à outils permettant d'obtenir les informations DILMO à partir de capteurs, et exploitent exclusivement du matériel spécifique. Le passage à l'échelle ne peut actuellement être réalisé en raison de la spécificité du matériel utilisé mais également en raison de l'absence d'approches de conception pour modéliser de tels environnements et les comportements proxémiques. Afin de faciliter l'intégration des capacités proxémiques dans les Interfaces Homme Machine, nous proposons une modélisation graphique basée sur un DSL permettant à des concepteurs non informaticiens d'exprimer les interactions proxémiques. Nous décrivons ici formellement notre proposition de DSL et nous l'illustrons par la mise en œuvre d'un prototype basé sur du matériel standard pour concevoir des comportements proxémiques dans différents environnements.

ABSTRACT. Proxemic interaction is an emerging area for improving HCI experiences. It describes how the five dimensions (i.e., Distance, Identity, Location, Movement, and Orientation -- DILMO) can be used to define interactions among people and digital devices. Current studies in this area are focused on developing proxemic applications with specific functionalities, based on toolkit that allow developers to obtain DILMO information from sensors. However, there exists a notable lack of general approaches able to support the whole implementation process, starting from the modeling of proxemic environments to represent general proxemics behaviors, and finalizing with the development of specific applications. In order to facilitate the integration of proxemic capabilities in Human Computer Interaction, we propose a graphical modelling based on Domain-Specific Language (DSL) that allows

designers to express proxemic interactions. We formally describe our proposed DSL and we applied it with a prototype to design proxemics behaviors in different proxemic environments.

Mots-clés : Interactions Proxémiques, , DSL, Modélisation, Conception,

KEYWORDS: Proxemic interaction, DSL, Modeling, Design.

1. Introduction

Le concept original de proxémique a été proposé par Edward T. Hall dans les années 60 (Hall, 1966). Il a présenté la façon dont les gens perçoivent, interprètent et utilisent l'espace, en particulier en ce qui concerne la distance physique entre les personnes (Evans *et al., 2000*). Ce concept a été repris par Greenberg en 2011 dans le contexte de l'Interaction Homme-Machine. Ainsi, l'interaction proxémique est un concept utilisé principalement pour décrire comment les personnes utilisent les distances interpersonnelles pour interagir avec les appareils numériques en termes de dimensions physiques proxémiques : Distance, Identité, Localisation, Mouvement et Orientation (DILMO) (Ballendat, 2010; Greenberg, 2011).

Certaines études se focalisent sur le développement d'applications proxémiques dans un cadre spécifique (Ballendat, 2010 ; Brock, 2018; Brudy, 2019; Cho, 2018; Evans, 2000; Garcia, 2019; Gørnb, 2019; Pérez, 2018; Sørensen, 2013). D'autres travaux proposent des cadres et une boîte à outils plus génériques qui aident les développeurs à traiter les informations proxémiques et à mettre en œuvre les comportements proxémiques des entités (Cadenas, 2017 ; Marquardt, 2013). Le comportement est ici à prendre comme une réaction face à une action d'une entité (*cf. 3.1- définition 1*). Cependant, il manque des approches plus génériques capables de soutenir l'ensemble du processus de réalisation des applications proxémiques pour les environnements intelligents, de la phase de conception au déploiement, en passant par le développement. De plus, il existe un besoin de standardisation, et dans un premier temps, le besoin de la définition d'une approche méthodologique intuitive et compatible avec du matériel standard (ie. les smartphones du marché)

Dans ce contexte, nous proposons une formalisation et un outil graphique de modélisation des interactions proxémiques pour des non spécialistes de la proxémique. Notre proposition est un DSL (Domain Specific Language) composé de symboles et de notations formelles. Les symboles représentent tous les composants d'un environnement proxémique, c'est-à-dire les entités avec lesquelles interagir, les dimensions DILMO et les comportements proxémiques (actions associées). À partir du modèle graphique, le DSL permet à des concepteurs non forcément informaticiens de créer un schéma XML. Ce schéma XML est ensuite utilisé pour générer automatiquement des classes Java qui seront incluses dans les applications pendant la phase de développement pour la collecte et l'enregistrement des informations proxémiques. Dans cet article, nous décrivons formellement cette modélisation graphique. Pour présenter l'utilisabilité du modèle graphique, nous développé un prototype du DSL permettant de concevoir les comportements proxémiques dans différents environnements et de générer les schémas XML.

La théorie sur la proxémique de Hall décrit les façons dont les individus perçoivent leur espace personnel en fonction de leur distance avec les autres (Gørnbk et O'Hara, 2016). Selon la théorie proxémique de Hall les zones d'interaction sont au nombre de 4 : intime (définie comme une distance entre 0 et 50 cm), personnelle (entre 0,5 et 1 m), sociale (entre 1 et 4 m) et publique (au-delà de 4 m). Il met en évidence le rôle des relations proxémiques comme éléments de communication entre les personnes. Les concepts de proxémique sont basés sur des facteurs physiques, sociaux et culturels qui influencent et régulent les interactions interpersonnelles (Marquard et al., 2011; 2013). Afin de savoir comment ces facteurs devrait être appliqués aux interactions proxémiques pour les applications informatiques, (Greemnerg et al., 2011) a identifié cinq dimensions : Distance, Identité, Localisation, Mouvement et Orientation (DILMO) qui sont associés à des personnes ou des appareils numériques. L'utilisation des interactions proxémiques a été exploitée par certains chercheurs pour concevoir des environnements intelligents basés sur l'interaction (Lendo, 2015; Rector 2004) et pour permettre aux utilisateurs de contrôler des appareils numériques de manière plus naturelle (Ballendat, 2010 ; Brudy, 2019 ; Lendo 2015; Marquard, 2013). L'interaction proxémique se situe au niveau de l'espace vital de l'homme et de la distance physique entre un individu et un dispositif numérique classé selon une échelle taxonomique proposée par (Brudy et al., 2019), tandis que les interfaces distribuées sont classées de manière dynamique dans la même taxonomie en raison des interactions entre les différentes plateformes numériques utilisées par les différents utilisateurs. Dans cette section, nous examinons quelques études récentes et remarquées dans ce domaine.

2.1. Méthode de classification des travaux

Afin de trouver, sélectionner et analyser des études centrées sur les outils pour soutenir la conception et la mise en œuvre d'applications proxémiques, nous avons fait une étude ciblée décomposée en trois étapes principales : (i) recherche des travaux relatifs au développement d'une application proxémique (ii) sélection des articles ; et (iii) élaboration d'un tableau comparatif basé sur l'ensemble de critères suivant :

1. **Type d'application :** cet aspect se rapporte à l'analyse du champ d'application, il permet d'évaluer s'il s'agit d'un outil permettant le développement d'applications ou d'une application particulière.
2. **Dimension proxémique :** cet aspect présente comment sont utilisées les dimensions proxémiques pour concevoir les environnements proxémiques.
3. **Méthodes de conception :** cet aspect évalue les méthodes et les modèles graphiques utilisés pour représenter les comportements proxémiques et concevoir des environnements proxémiques.
4. **L'utilisabilité :** cet aspect spécifie si les travaux ont offert des solutions pour concevoir les interactions et environnements proxémiques ainsi que la disponibilité du code source.

2.2. Analyse comparative

Le Tableau 1 résume l'analyse qualitative et comparative que nous décrivons ci-dessous.

La plupart des travaux portent sur la mise en œuvre d'applications pour des fonctions spécifiques. Des études récentes comme (Ballendat, 2010; Brudy 2018, Gørnb, 2018, Lendo) concernent le contrôle des appareils numériques dans des environnements intelligents pour mettre en œuvre des lecteurs de médias interactifs, de l'analyse vidéo et des applications de contrôle à distance. Des applications proxémiques ont également été utilisées pour soutenir l'interaction gestuelle et améliorer la vie quotidienne des personnes aveugles (Brock, 2018 ; Dingler 2009 ; Garcia 2019; Mentis, 2012; Mojgan, 2018; Rector, 2017). Des applications mobiles proxémiques ont été développées grâce aux nouvelles capacités des appareils mobiles qui permettent aux développeurs d'obtenir des informations proxémiques à partir des capteurs des smartphones (Brock, 2018 ; Cho, 2018; Perez, 2018; Sørensen, 2013). Certains travaux développent des applications proxémiques pour adapter les visuels sur les écrans (Dostal, 2014; Vermulen, 2015; Wolf, 2016). La manière dont les interactions proxémiques peuvent être utilisées pour les robots de service est décrite dans (Bhagya et al., 2018). Seuls quelques travaux ont développé des frameworks et des outils pour construire et concevoir des applications basées sur des interactions proxémiques (Cardenas, 2017; Gørnbk, 2016; Kim 2016; Marquard 2011). La majorité des travaux utilise tout ou partie des dimensions DILMO. Les environnements proxémiques basés sur DILMO, dans lesquels toutes les dimensions sont prises en compte, sont décrits dans (Ballendat 2010; Bhagya, 2018; Brudy, 2019; Cardenas, 2017; Dostal, 2014; Garcia, 2019; Kim 2016; Ledo, 2015; Marquardt, 2011; Mentis, 2012; Mojgan, 2018; Rector, 2017). Le travail présenté dans (Cardenas et Garcia, 2017) illustre comment les dimensions proxémiques peuvent soutenir l'interaction entre entités (personnes et objets) en tenant compte du contexte. L'étude présentée dans (Dostal et al., 2014) montre comment les personnes peuvent interagir dans différences zones et distances autour d'un dispositif spécifique.

Les environnements proxémiques réduits aux dimensions DIMO sont pris en compte dans les études présentées dans (Brock, 2018 ; Wang 2012). Dans (Brock, 2018) est présentée une application capable d'ajuster le contenu de l'affichage d'un dispositif identifié en fonction de la distance, de la localisation, du mouvement et de l'orientation des utilisateurs. Dans le contexte des jeux, trois environnements proxémiques DLMO sont présentés dans (Gørnbk et al.,2019). Les actions des jeux sont établies en fonction de la distance, de la localisation, des mouvements et de l'orientation des appareils manipulés par les enfants. Dans (Cho et al.,2018), un environnement proxémique DIL est décrit dans le contexte d'une application qui permet la reconnaissance de matériaux grâce à une caméra thermique mobile à bas prix intégrée dans un smartphone. Cette application mesure la distance physique entre la caméra et le matériau et reconnaît une texture spécifique (l'identité). Des travaux basés sur les environnements DILM sont présentés dans (Dingler, 2015; Sørensen. 2013). Par exemple, le système de musique multi-pièces proposé dans

(Sørensen *et al*., 2013) est une application mobile basée sur des interactions proxémiques et permet à l'utilisateur d'écouter la même playlist malgré ses déplacements grâce à l'activation de différents haut-parleurs disposés dans la maison. Les environnements proxémiques réagissant à la distance, à l'identité et à l'orientation (c'est-à-dire les environnements proxémiques DIO) sont présentés dans (Dostal, 2013; Gørnbk, 2016; Wolf, 2016). L'étude présentée dans (Gørnbk *et al*., 2016) détaille l'utilisation de la boussole intégrée dans les appareils mobiles qui a permis d'identifier et soutenir le processus d'appariement basé sur la distance et l'orientation. Dans (Perez *et al*., 2018), un environnement proxémique DIM est décrit dans le contexte d'une application mobile de premiers secours (FAMA). FAMA offre aux sauveteurs la possibilité d'obtenir l'identité d'une personne inconsciente en cas d'urgence lorsqu'ils se déplacent dans les zones proxémiques de la personne blessée. La distance et la localisation (DL) des personnes sont prises en compte dans (Vermeu et al, 2015) pour la définition d'actions dans un affichage interactif.

En ce qui concerne les méthodes de conception, seuls trois travaux ont présenté des outils pour concevoir les interactions proxémiques. SpiderEyes (Dostal et al. 2014) est un système qui fournit un outil visuel qui permet aux développeurs de concevoir des applications proxémiques collaboratives. Le travail présenté dans (Kim et al., 2016) illustre les interactions proxémiques dans des environnements intelligents en utilisant des modèles miniatures. Cependant, il faut du matériel spécifique pour concevoir ces environnements tels que des marqueurs de réalité augmentée (AR) et un projecteur de caméra. Dans (Marquardt et al., 2011) est présenté une boîte à outils composée d'une collection de bibliothèques en C et d'une architecture à composants qui prend en compte les informations spatiales et les relations entre les objets et l'espace. Toutefois, il n'offre pas de notation graphique pour permettre aux utilisateurs finaux la modélisation des comportements proxémiques.

Tableau 1. L'évaluation comparative

Author	Type	D	I	L	M	O	Méthodes de conception	L'utilisabilité	
								Conception intuitive	Disponibilité du code source
(Ballendat et al.,2010)	Application	■	■	■	■	■			
(Bhagya et al.,2018)	Service robot	■	■	■	■	■			
(Brock et al.,2018)	Application	■	■		■	■			
Brudy, 2019	Application	■	■	■	■	■			
(Cardenas et al.,2017)	Framework	■	■	■	■	■			■
(Cho et al.,2018)	Application	■	■	■					■
(Dingler et al.,2015)	Application	■	■	■	■				

(Dostal et al.,2014	Application	■	■	■	■	■	■		
(Dostal et al.,2013)	Application	■	■			■			
(Garcia et al.,2019)	Application	■	■	■	■	■			
(Gørnbk et al.,2019)	Application	■		■	■	■			
(Gørnbaek et al.,2016)	Application	■	■			■			
(Kim et al.,2016)	Application	■	■	■	■	■	■	■	
(Ledo et al.,2015)	Application	■	■	■	■	■			
(Marquardt et al.,2011)	Framework	■	■	■	■	■	■		■
(Mentis et al., 2012)	Application	■	■	■	■	■			
(Mojgan et al. 2018)	Application	■	■	■	■	■			
(Perez et al.,2018)	Application	■	■		■				
(Rector et al.,2017)	Application	■	■	■	■	■			
(Sørensen et al.,2013)	Application	■	■	■	■				
(Vermeu et al., 2015)	Application	■		■					
(Wang et al.,2012)	Framework	■	■		■	■			
(Wolf et al. ,2016)	Application	■	■			■			

L'utilisabilité est le critère le moins rencontré dans la grande majorité des propositions antérieures. Nous n'en avons trouvé qu'un qui offre des notations intuitives pour concevoir facilement des applications proxémiques (Kim et al, 2016). Ce cadre fournit des outils qui aident les développeurs à construire une application frontale. Cependant, il ne fournit pas de notation graphique pour permettre au concepteur de modéliser les comportements proxémiques.

3. Modélisation du comportement proxémique

Dans cette section, nous présentons une description générale de la modélisation du comportement proxémique fondée sur un DSL. Le DSL permet la représentation des environnements et des comportements proxémiques à partir de conditions de départ spécifiques des objets et entités physiques basées sur les dimensions DILMO. À partir de la modélisation graphique d'un scénario proxémique, un schéma XML est généré, afin de faciliter le développement d'applications proxémiques. La modélisation et la standardisation permettent que les concepteurs aient un langage commun pour exprimer les comportements proxémiques de chaque entité dans un environnement proxémique.

3.1. Définitions formelles pour les environnements proxémiques

Un environnement proxémique (défini précisément plus bas) est décrit à partir d'une entité cible (CPS, définie plus bas) qui réagit selon les dimensions DILMO avec toutes les autres entités. La première définition concerne donc les entités qui peuvent interagir dans un environnement proxémique

Définition 1. **Entité (E),** une entité, désignée par E, représente un objet d'interaction telle qu'une personne, un objet ou un dispositif, qui peut être identifié ou non, dans un espace physique.

Définition 2. **Cyber Physical System (CPS)**, représente l'entité cible, à partir de laquelle un environnement proxémique (E_P) est défini. Toutes les zones proxémiques (Z_P) et les dimensions DILMO sont mesurées pour toutes les entités (E) par rapport au CPS.

Les définitions 3 à 6 portent sur les attributs proxémique.

Définition 3. L'identité, notée (I), représente une entité (E) clairement identifiée dans l'environnement proxémique. Par analogie avec la définition 1, il s'agit de 'cette' personne, 'ce' dispositif ou de 'cet' objet.

Définition 4. **La distance, notée (D)**, est la mesure physique entre des entités.

Définition 5. **Localisation**, **notée (L),** représente les coordonnées cartésiennes indiquant la localisation d'une entité (E) par rapport au CPS dans chaque zone proxémique ; elle est désignée comme E.L = (x_1, y_1) ou I.L = (x_1, y_1).

Définition 6. **Le mouvement, noté (M),** est un vecteur représentant la magnitude et la direction d'une entité (E ou I) par rapport au CPS. Elle est désignée par E.M = (m,d) ou I.M = (m,d) ;

- Si d = 1, E ou I se déplace vers le CPS.
- Si d = 0, E ou I est statique
- Si d= -1, E ou I s'éloigne du CPS.

Définition 7. **L'orientation, notée (O)** représente l'angle relatif à l'alignement face à face des entités par rapport au CPS, elle est désignée comme E.O ou I.O.

L'environnement proxémique est l'espace physique dans lequel les entités peuvent interagir avec le CPS selon les variables DILMO.

**Définition 8. Zone proxémique (Z_P*),* une zone proxémique représente une zone circulaire délimitée par une distance maximale par rapport au CPS. Il existe quatre zones proxémiques définies en fonction de cette distance maximale *: $Z\_P_{intimité}$, $Z\_P_{personnelle}$, $Z\_P_{sociale}$, et $Z\_P_{publique}$.*

La zone proxémique dans laquelle une entité est située est désignée comme : E.P_ Z ou I.Z_P. Elle peut prendre les valeurs suivantes : tt

- $E.Z\_P = Z\_P_{intimité}, si\ \ 0 < E.D \leq MAX\_ID;$
- $E.Z\_P = Z\_P_{personnelle}, si\ \ MAX\_ID < E.D \leq MAX\_PD;$
- $E.Z\_P = P\_Z_{sociale}, si\ MAX\_PD < E.D \leq MAX\_SD;$
- $E.Z\_P = Z\_P_{publique}, si\ \ MAX\_SD < E.D \leq MAX\_PubD;$

où MAX_ID, MAX_PD, MAX_SD, et MAX Pub_D représentent les distances maximales définissant respectivement les zones intimes, privées, sociales, et publiques. Il s'agit de paramètres qui doivent être fournis par les utilisateurs ou les développeurs. Par analogie, la même notation s'applique pour les entités identifiées (I) I.Z_P.

Définition 9. **Environnement Proxémique, noté (E_P)**, représente un ensemble de capteurs et de dispositifs attachés à des entités (E) qui peuvent à leur tour interagir selon (DILMO) dans le cadre d'un scénario.

Définition 10. Comportement (B), un comportement représente le changement des mesures D, L, M et O ou du Z_P d'une entité (E ou I), à partir de son comportement initial (B_0), sachant que l'identité de I ne changera pas. Le comportement est désigné par un tuple $E.B_i = <E.D_i, E.L_i, E.M_i, E.O_i, E.Z\_P_i>$

Cette notation est produite par une transition à partir du comportement précédent E_{i-1}, de telle sorte que :

$$E.B_0 = < E.D_0, E.L_0, E.M_0, E.O_0, E.Z\_P_0 > \text{ (le comportement initial)}$$

- ∇ = représente le changement des conditions d'une entité, il existe une fonction définie par l'utilisateur pour détecter le changement de dimensions DLMO
- $E.B_i = \nabla\ E.B_{i-1} = <E.D_i, E.L_i, E.M_i, E.O_i, E.Z\_P_i>$ (for $i \geq 1$)

En fonction du comportement d'une entité (E) dans un E_P il est possible de détecter le changement d'une mesure DLMO qui provoquera automatiquement le changement d'autres mesures DLMO et de sa zone proxémique. Par exemple, si un mouvement (M) de E est détecté, cela causera le changement de sa localisation (L) et de sa distance (D) ; De même, si la distance (D) de E change, cela signifie que E se déplace et que sa localisation (L) changera également. Dans ces deux exemples, il est également possible que la Z_P de E change.

Définition 11. Action (Action), une Action représente un événement réalisé par le CPS ou toute autre entité (E ou I), dans un E_P, en réponse à un comportement spécifique (B) d'une entité ou d'un groupe d'entités.

Pour définir un E_P, les conditions sont donc :

1. Distances, définies pour les quatre Z_P fournies par les concepteurs/développeurs en fonction des besoins des utilisateurs.
2. Un E_P existe seulement dans le cas , où il contient une entité cible (CPS), des entités et des distances :

   E_P = < CPS, entités, identités, distances >, le CPS n'est pas NULL

3. Le CPS représente la cible d'interaction et a les propriétés suivantes :
   - la localisation d'origine du CPS est indiquée comme CPS.L (0,0), puisque les zones proxémiques et les dimensions DILMO de tous les autres objets d'interaction sont déterminées par rapport au CPS ;
   - Le champ de vision est défini selon les angles suivants : ∡ MinAofV et ∡ MaxAofV. Ces valeurs sont des paramètres fournis par les utilisateurs ou les développeurs.
4. Pour toutes les entités (E) du système, les dimensions DILMO et les zones proxémiques peuvent avoir des conditions initiales :

   $\forall E_i \in E\_P \wedge I_j \in E\_P$, $E_i.B_0 = < D_0,L_0,M_0,O_0 > \wedge I_j.B_0 = <D_0,I_j,L_0,M_0,O_0>$ et L_0,M_0,O_0 peuvent être NULLE.

Cela signifie que la distance (D) est la seule dimension obligatoire pour les entités. L'orientation est une dimension non pertinente si le CPS n'a pas de champs de vision c'est à dire que ∡ MinAofV et ∡ MaxAofV sont égaux à zéro. Par exemple si nous définissons une action qui est basée sur la distance entre les appareils avec l'utilisation du Bluetooth seulement.

Ces définitions forment la base formelle de notre DSL. Elles sont utilisées pour spécifier l'E_P, avec ses entités et leurs actions, qui sont définies à partir de conditions initiales (Z_P et DILMO).

3.2. Notation graphique, à base de symboles, pour les environnements proxémiques

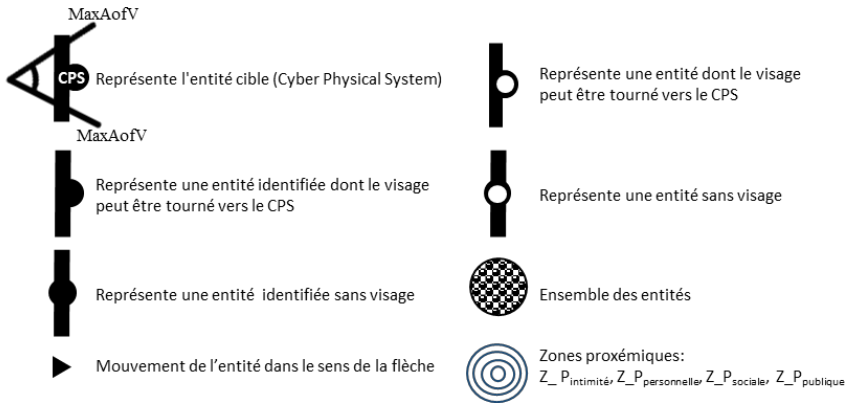


Figure 1. Symboles de la notation Graphique pour les env. proxémiques

Notre modèle graphique s'inspire du système de notation proposé par Edward T. Hall dans (Hall, 1963; 1966) qui fournit un code de notation permettant la représentation des relations humaines. La figure 1 décrit les éléments graphiques que nous proposons. Nous illustrons le formalisme de la notation graphique sur l'exemple suivant :

Soit un E_P, avec les identités I_1 et I_2 se dirigeant vers le CPS depuis la zone personnelle ($Z\_P_{personnel}$) et faisant face au CPS et l'entité E_1 se tenant devant le CPS dans la zone sociale ($Z\_P_{sociale}$). La figure 2 montre la représentation graphique de ce scénario avec les conditions initiales suivantes :

- Les quatre tuples de E_P: E_P =< CPS, entités = $\{E_1\}$, identités = $\{I_1,I_2\}$, distances = {0,4,1,4,7} >
- CPS.L = (0,0) ; ∡ MinAofV = 0^0 et MaxAofV = 60^0;
- Distance des entités ; I_1.D = 0.85, I_2.D = 0.80, E_1.D = 3 ;
- Localisation de l'I1.L = (0.8,0.3) ;
- Mouvement de l'I_1 et l'I_2 (vers CPS) : I_1.M(l_1,1) et I_2.M(l_2,1) ; Mouvement de l'E_1 (statique) : E_1.M=(l_1,0) ;
- Orientation de l'entités (les visages des entités sont dans le champs de vision du CPS): I_1.O=20^0, I_2.O=35^0, E_1.O= 75^0 ainsi : ∡ MinAofV∡ I_1.O,E_1.O ∡ MaxAofV;
- Zones proxémiques des entités $I_1.Z\_P$, $I_2.Z\_P = Z\_P_{personnelle}$, $E_1.P = Z\_P_{sociale}$;
- Ainsi, les comportements initiaux des entités sont :
- $I1.B_0$ = <0.85, (0.8,0.3), (l_1,1), 20^0, $Z\_P_{personnelle}$>
- $I2.B_0$ = <0.8,NULL, (l_2,1),35◦, $Z\_P_{personnelle}$>
- $E1.B_0$ = <3, NULL, (l_3,0),75◦, $Z\_P_{sociale}$>

À partir de ces conditions initiales, le CPS peut réagir par des actions qui peuvent être désignées de plusieurs façons, par exemple : $Z\_P_{sociale}$

Si ∃ E dans $Z\_P_{sociale}$, alors $CPS.Action_1$.

∀ I ∈ P E.identités ∧ I_j dans $Z\_P_{personnelle}$ puis $CPS.Action_1$. Ainsi, selon les comportements des entités détectées (E ou I), les actions du CPS peuvent changer. Elles sont représentées par des notations formelles.

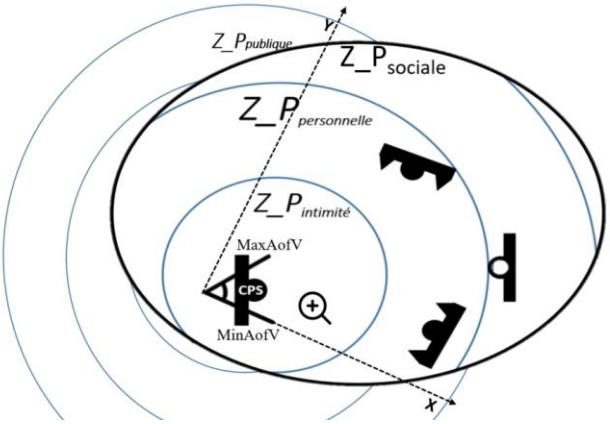


Figure 2. Modèle graphique d'un E_P

Une fois les comportements modélisés dans ce scénario (concepteur), les actions doivent être implémentées (développeur). Cette étape nécessite la production de code. De même, la manière dont les comportements peuvent être détectés en fonction de la technologie utilisée dans le E_P (par exemple, BLE Bluetooth à base consommation, capacités des capteurs de l'appareil, vision par ordinateur) nécessitent l'intervention et la production de code. Lorsqu'un E_P possède plusieurs CPS, il sera nécessaire d'itérer la modélisation pour chaque CPS.

Le choix entre E et I dépend des entités manipulées, des actions à réaliser et de leur action générique. Prenons un scénario simple : lorsque deux personnes (E) s'approchent d'un panneau publicitaire (CPS), il s'allume et une publicité est affichée.

En remplaçant E par I et ainsi pouvant connaître les relations entre les deux I, il est possible de cibler et différencier la publicité lorsqu'il s'agit de couples ou d'amis par exemple.

La Figure 3 présente en résumé le diagramme de classe correspondant :

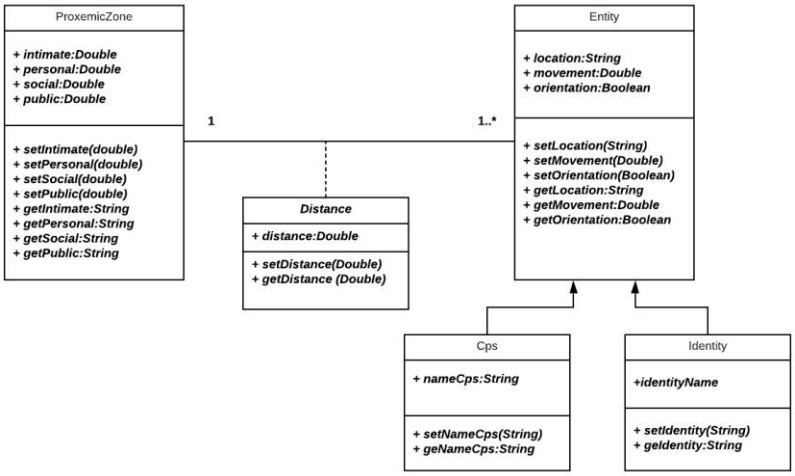


Figure 3 : Diagramme de classes

Dans la section suivante, nous verrons comment se traduit la représentation formelle en un fichier XML de type Schéma Design (XSD).

3.3. XSD pour les environnements proxémiques

A partir des spécifications formelles et de la conception graphique d'un environnement proxémique, il est possible de générer automatiquement un fichier XML de type Schéma Design (XSD). Il contient les informations relatives aux conditions initiales d'un E_P, les dimensions DILMO, les actions et les zones. Le XSD permet de traduire le modèle de conception dans un langage informatique (Daum,2003 ; Machkour,2016). Nous utilisons ensuite le générateur de classes JAXB (Oracle, 2019) qui permet de créer des interfaces et d'implémenter des classes à partir de cet XSD. Notre proposition de DSL graphique constitue un outil qui supporte à la fois les phases de conception et de développement des applications proxémiques.

- Conception de l'environnement proxémique : Cela signifie que les zones proxémiques (Z_P), le CPS, les entités (E et I), l'ensemble des dimensions DILMO à prendre en compte et les conditions initiales sont identifiées. A partir de la définition formelle, il est possible d'établir les règles d'inférence qui peuvent générer des comportements spécifiques et des actions liées aux déplacements des entités dans l'environnement proxémique.
- Modélisation de l'environnement proxémique : la notation graphique E_P est représentée dans un modèle graphique intuitif.

- Modélisation des comportements proxémiques : une fois l'E_P défini et modélisé, il est possible de modéliser les comportements proxémiques dans le temps en fonction des dimensions DILMO considérées pour représenter les actions proxémiques du CPS et des entités.
- Génération du XSD : les informations relatives aux conditions initiales d'un E_P ainsi que les comportements et actions proxémiques sont stockées dans un fichier XML. Le XSD représente chaque composant de l'environnement proxémique tel que le concepteur l'a créé. Les interfaces et classes Java sont ensuite automatiquement générées à partir du XSD. Cette génération de code est basée sur des règles one-to-one (une entité génère une classe, une propriété génère un attribut).
- Implémentation d'une application proxémique : les interfaces et classes générées à l'étape précédente peuvent ensuite être intégrées intégrées, et utilisées. Dans la phase suivante de mise en œuvre, les développeurs décident comment les dimensions DILMO sont capturées en fonction de la technologie des capteurs disponibles (par exemple, BLE, caméra, , boussole, etc.) à l'aide de l'API développées et accessible sur : https://github.com/paulocpd76/DSLGUI003-L.

4. Un premier prototype du DSL

La figure 4 montre la représentation des conditions initiales de l'E_P décrites dans la section 3.2 à l'aide de notre DSL graphique. Le concepteur précise les distances qui définissent les quatre zones proxémiques, ajoute des entités ou des identités sur chaque zone (représentées dans les carrés noirs), ajuste les dimensions DILMO dans l'icône qui représente chaque entité (dans la zone verte), et spécifie qu'une action est effectuée par le CPS (en appuyant sur le bouton « action »). Le curseur au-dessus de chaque entité permet de simuler les changements de distance entre l'entité et le CPS. D'autres comportements d'entités sont simulés en modifiant leurs dimensions DILMO sur leur icône respective dans la zone verte. Les fichiers XSD et XML sont générés automatiquement pour stocker, lire et modifier les conditions initiales, ainsi que tous les autres comportements simulés dans l'interface graphique. Ce prototype est librement téléchargeable sur github mentionné précédemment.

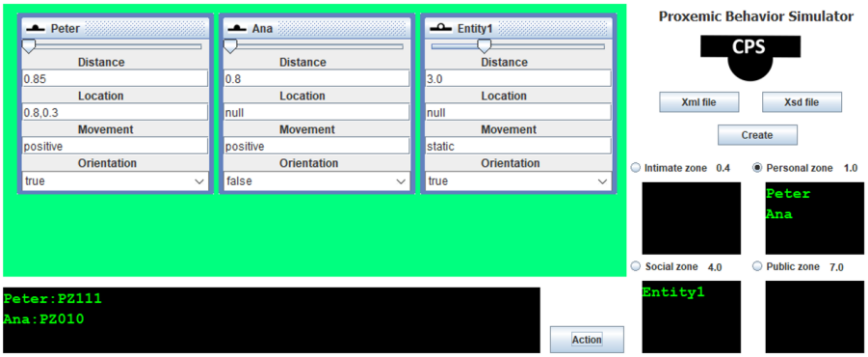


Figure 4 : Modélisation d'un E_P avec DSL

Il est possible de télécharger différents fichiers XML et XSD afin de réaliser des simulations et afin de définir les comportements et les actions souhaités dans un E_P (permet de s'affranchir de la phase de conception interactive). Dans ce premier prototype, l'orientation des entités est représentée comme un paramètre booléen indiquant si l'entité est ou non dans le champ de vision du CPS. Pour la prochaine version, il est prévu d'améliorer et d'affiner cette fonctionnalité en indiquant les angles dans la zone de vision du CPS et l'orientation des entités. Nous prévoyons également d'ajouter tous les éléments graphiques présentés à la section 3.2 et de permettre leur utilisation par simple glisser/déposer.

5. Conclusion

Dans cet article, nous avons présenté une modélisation graphique basée sur un DSL permettant de mieux appréhender et de mieux comprendre les environnements proxémiques. Contrairement aux approches plus traditionnelles des règles de type ECA (Event-Condition-Action) centrée sur l'occurrence d'un événement dans une perspective ensembliste, la conception est ici centrée sur des entités (ici proxémiques), ce qui simplifie le travail d'identification des conditions de déclenchement mais également rend la conception plus intuitive, quitte à la dupliquer pour chaque CPS. A l'aide d'un fichier XSD (fichier XML de type Schéma Design) généré par le DSL, nous générons des classes afin de faciliter le travail in fine des développeurs pour rendre plus rapide et plus simple l'intégration dans le code de la capture des données via les capteurs, et les actions associées aux événements. Ce travail est grandement facilité par l'utilisation d'une API (Perez, 2018), fonctionnant pour dispositifs mobiles équipés du système Android Nous avons décliné la modélisation graphique en une interface tangible qui incarne les données symboliques par des objets physiques. Cette interface semble naturellement faciliter la collaboration au moment de la conception, par l'appropriation des mouvements dans l'espace, mais aussi la compréhension des symboles grâce à leur incarnation. Nos travaux futurs devront montrer s'il y a un réel apport à la physicalisation du DSL. La méthode de conception et son adéquation à l'API sont (au moment de l'écriture de cet article) en train d'être testées par des élèves ingénieurs et par les développeurs de l'entreprise Dev 1.0, offrant ainsi une évaluation in-vivo du concept et des outils proposés.

Remerciements

Travaux soutenus par la région Nouvelle Aquitaine et la Communauté Pays Basque dans le cadre du projet PISCO, réalisé en collaboration par l'UPPA, l'ESTIA INSTITUTE OF TECHNOLOGY et la société Dev 1.0.

Bibliographie

Ballendat, T., Marquardt, N., Saul, G. (2010). Proxemic interaction: designing for a proximity and orientation-aware environment. In acm International Conference on Interactive Tabletops and Surfaces, p. 121-130

Bhagya, S., Samarakoon, P., Sirithunge, H.C., Viraj, M., Muthugala, J., Buddhika, A., Jayasekara, P. (2018). Proxemics and approach evaluation by service robot based on user behavior in domestic environment. International Conference on Intelligent Robots and Systems (IROS). IEEE. p. 8192-8199

Brock, M., Quigley, A., Kristensson, P.O. (2018). Change blindness in proximity-aware mobile interfaces. CHI conf. on Human Factors in Computing Systems. p. 1-7

Brudy, F., Holz, C., Radle, R., Wu, C.J., Houben, S., Klokmose, C.N., Marquardt, N. (2019). Cross-device taxonomy: Survey, opportunities and challenges of interactions spanning across multiple devices. CHI conf. on Human Factors in Computing Systems, p. 1-28

Brudy, F., Suwanwatcharachat, S., Zhang, W., Houben, S., Marquardt, N. (2018). Eagleview: A video analysis tool for visualising and querying spatial interactions of people and devices. International conf. on Interactive Surfaces and Spaces. p.61-72

Cardenas, C., Garcia-Macias, J.A. (2017). Proximithings: Implementing proxemic interactions in the internet of things. Procedia Computer Science vol, 113, p. 49-56.

Cho, Y., Bianchi-Berthouze, N., Marquardt, N., Julier, S.J. (2018). Deep thermal imaging: Proximate material type recognition in the wild through deep learning of spatial surface temperature patterns. CHI conf. on Human Factors in Computing Systems. p. 1-13

Daum, B. (2003) Modeling business objects with XML schema. Morgan Kaufmann, San Francisco, Calif. Heidelberg, Germany

Dingler, T., Funk, M., Alt, F. (2015). Interaction proxemics: Combining physical spaces for seamless gesture interaction. In Proc of Internat Symposium on Pervasive Displays, vol.107, p. 114

Dostal, J., Hinrichs, U., Kristensson, P.O., Quigley, A. (2014). Spidereyes: designing attention-and proximity-aware collaborative interfaces for wall-sized displays. International conf. on Intelligent User Interfaces, p. 143-152

Dostal, J., Kristensson, P.O., Quigley, A. (2013). Multi-view proxemics: distance and position sensitive interaction. In the acm Internat Symposium on Pervasive Displays, p. 1-6.

Evans, G.W., Lepore, S.J., Allen, K.M. (2000), Cross-cultural differences in tolerance for crowding: Fact or fiction? Journal of Personality and Social Psychology, vol. 79, p. 204.

Garcia-Macias, J.A., Ramos, A.G., Hasimoto-Beltran, R., Hernandez, S.E.P. (2019): Uasisi: a modular and adaptable wearable system to assist the visually impaired. Procedia Computer Science, vol. 151, p. 425-430.

Greenberg, S., Marquardt, N., Ballendat, T., Diaz-Marino, R., Wang, M. (2011). Proxemic interactions: the new ubicomp? Interactions, vol. 18, p. 42-50.

Gørnbk, J.E., Linding, C., Kromann, A., Jensen, T.F.H., Petersen, M.G. (2019). Proxemics play: Exploring the interplay between mobile devices and interiors. In Companion Publication on Designing Interactive Systems Conference, p. 177-181

Gørnbk, J.E., O'Hara, K. (2016). Built-in device orientation sensors for ad-hoc pairing and spatial awareness. In Proc. of Cross-Surface Workshop

Hall, E.T. (1963). A system for the notation of proxemic behavior. American Anthropologist, vol. 65, p.1003-1026.

Hall, E.T. (1966). The Hidden Dimension: An anthropologist examines man's use of space in private and public. New York: Anchor Books; Doubleday & Company, Inc.

Kim, H.J., Kim, J.W., Nam, T.J. (2016). ministudio: Designers' tool for prototyping ubicomp space with interactive miniature. CHI conf. on Human Factors in Comp. Syst, p. 213-224.

Ledo, D., Greenberg, S., Marquardt, N., Boring, S. (2015). Proxemic-aware controls: Designing remote controls for ubiquitous computing ecologies. In Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services, p. 187-198

Machkour, M., Afdel, K. (2016). Transforming xml into object-relational schema. IOSR Journal of Computer Engineering, vol. 18, p. 40-52.

Marquardt, N. (2013). Proxemic interactions in ubiquitous computing ecologies. Ph.D. thesis, University of Calgary.

Marquardt, N., Diaz-Marino, R., Boring, S., Greenberg, S. (2011). The proximity toolkit: prototyping proxemic interactions in ubiquitous computing ecologies. In Proceedings of the 24th annual ACM symposium on User interface software and technology, p. 315-326.

Mentis, H.M., O'Hara, K., Sellen, A., Trivedi, R. (2012). Interaction proxemics and image use in neurosurgery. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, p. 927-936.

Mojgan, G., Marvin, P., Wong, C., Wallace, J.R., Scott, S.D. (2018). Increasing passersby engagement with public large interactive displays: A study of proxemics and conation. International Conference on Interactive Surfaces and Spaces, p. 1-14.

Oracle: Using the jaxb class generator and jaxb users guide (2019) https://docs.oracle.com/javase/8/docs/technotes/guides/xml/jaxb/index.html

Perez, P., Roose, P., Marc, D., Couture, N., Cardinale, Y., Masson, D. (2018). Proxemics for first aid to unconscious injured person. In Proceedings of the 30th Conference on l'Interaction Homme-Machine, p. 156-162

Rector, K., Salmon, K., Thornton, D., Joshi, N., Morris, M.R. (2017). Eyes-free art: Exploring proxemic audio interfaces for blind and low vision art engagement. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, vol, 1, p. 1-21

Roussel, N., Evans, H., Hansen, H. (2004). Mirrorspace: using proximity as an interface to video-mediated communication. In Int'l Conference on Pervasive Computing, p. 345-350.

Sørensen, H., Kristensen, M.G., Kjeldskov, J., Skov, M.B. (2013). Proxemic interaction in a multi-room music system. In Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration, p. 153-162.

Vermeulen, J., Luyten, K., Coninx, K., Marquardt, N., Bird, J. (2015). Proxemic flow: Dynamic peripheral floor visualizations for revealing and mediating large surface interactions. In IFIP Conference on Human-Computer Interaction, p. 107-114.

Wang, M., Boring, S., Greenberg, S. (2012). International symposium on pervasive displays, p. 1-7.

Wolf, K., Abdelrahman, Y., Kubitza, T., Schmidt, A. (2016). Proxemic zones of exhibits and their manipulation using floor projection. In Proceedings of the 5th acm International Symposium on Pervasive Displays, p. 33-37.

Wolf, K., Lischke, L. (2014). Urban proxemics for public guidance In Proceedings of the NordiCHI Workshop on Ubicomp beyond Devices: People, Objects, Space and Meaning, acm, vol.10 p. 2639189-2654842.

Xatkit: A model-based chatbot development framework - Extended Abstract

Gwendal Daniel [1], **Jordi Cabot** [1,2], **Laurent Deruelle** [3], **Mustapha Derras** [3]

1. IN3, Universitat Oberta de Catalunya (UOC), Spain
gdaniel@uoc.edu

2. ICREA, Spain
jordi.cabot@icrea.cat

3. Berger-Levrault, France
[laurent.deruelle,mustapha.derras]@berger-levrault.com

KEYWORDS: *chatbot, bot, low-code, DSL*

Instant messaging platforms have been widely adopted as one of the main technologies to communicate and exchange information. Nowadays, most of them provide built-in support for integrating *chatbot applications*, which are automated conversational agents capable of interacting with users of the platform. Chatbots have proven useful in various contexts to automate tasks and improve the user experience, such as automated customer services, education, and e-commerce. Chatbot design will become a key ability in IT hires in the near future.

This widespread demand has emphasized the need to be able to quickly build non-trivial chatbot applications supporting AI-based natural language processing (NLP) and being capable of taking part in orchestrations (Brambilla *et al.*, 2009) of internal and external services to carry out user requests. As such, chatbots are becoming complex software artifacts that require a methodical development approach encompassing a variety of technical domains, ranging from NLP to a deep understanding of the APIs of the targeted instant messaging platforms and third-party services. So far, chatbot development platforms have mainly addressed the NLP challenge, typically by relying on external *intent recognition providers*, which are NLP frameworks providing user-friendly interfaces to define conversation assets. As a trade-off, chatbot applications are tightly coupled to those providers, hampering their maintainability and reusability. Similarly, current chatbot platforms lack proper abstraction mechanisms to easily integrate and communicate with other external platforms the bot needs to interact with.

We aim to tackle all these issues by raising the level of abstraction at what chatbots are defined. To this purpose, we introduce Xatkit (Daniel *et al*., 2020), a novel model-based chatbot development framework that aims to address this question using Model Driven Engineering (Brambilla *et al*., 2017) techniques: domain-specific languages, platform independent bot definitions, and runtime interpretation. Indeed, Xatkit embeds a dedicated chatbot-specific modeling language to specify user intentions, computable actions and callable services, combining them in rich conversation flows.

The resulting chatbot definition is independent of the intent recognition provider (which can be configured as part of the available Xatkit options) and frees the designer from the technical complexities of dealing with messaging and backend platforms as Xatkit can be deployed through the Xatkit runtime component on them without performing any additional steps. The Xatkit framework is open source[1].

Xatkit is ready to be used in real-case scenarios. But it has still plenty of room for improvements. At the language level we plan to improve the variability of the bot specification, moving towards a product-line approach that enables companies to create and quickly update several versions of the same bot (e.g. to create localized versions of the bot for each branch of the company). At the framework level, we plan to work on the integration of chatbot generators, able to create partial bot specifications from existing data sources within the company (e.g. FAQs or user guides). We also plan to study the combination of sentiment analysis and behavioural design patterns (Fogg, 2002) to create more likeable and effective chatbots (Ren *et al*., 2019). Finally, security and access-control is another important aspect of any chatbot design as we may want to allow users to query (or not) certain aspects of our data depending on their profile.

References

Brambilla M., Cabot J., Wimmer M. (2017). *Model-driven software engineering in practice, second edition*. Morgan & Claypool Publishers.

Brambilla M., Dosmi M., Fraternali P. (2009). Model-driven engineering of service orchestrations. In *2009 IEEE congress on services, Part I, SERVICES I 2009*, pp. 562–569.

Daniel G., Cabot J., Deruelle L., Derras M. (2020). Xatkit: A multimodal low-code chatbot development framework. *IEEE Access*, Vol. 8, pp. 15332–15346.

Fogg B. J. (2002, December). Persuasive technology: Using computers to change what we think and do. *Ubiquity*, Vol. 2002, No. December.

Ren R., Castro J. W., Acuña S. T., Lara J. de. (2019). Usability of chatbots: A systematic mapping study. In *The 31st int. conf. on software engineering and knowledge engineering, SEKE 2019.*, pp. 479–617.

1. https://github.com/xatkit-bot-platform

Aide à la décision et recommandation

Marketing des traces : du tracking, des contre-mesures et de leur efficacité - *Robert Viseur* (article long)

Quelle Blockchain choisir ? Un outil d'aide à la décision pour guider le choix de technologie Blockchain - *Nicolas Six, Nicolas Herbaut et Camille Salinesi* (article long)

Recommandations basées sur les centres d'intérêts utilisateurs en Business Intelligence - *Krista Drushku, Julien Aligon, Nicolas Labroche, Patrick Marcel and Verónika Peralta* (résumé étendu)

Marketing des traces : du *tracking*, des contre-mesures et de leur efficacité

Robert Viseur[1]

1. FWEG - Universite de Mons
Service de Technologies de l'Information et de la Communication
17, place Warocqué, B-7000 Mons
robert.viseur@umons.ac.be

RÉSUMÉ. D'un Web de documents à l'intérêt commercial incertain, porté par des pionniers croyant au partage des connaissances, le Web a par la suite évolué vers une forme collaborative et temps réel rentabilisée par la publicité. Cette dernière a évoluée vers la publicité ciblée incluant la publicité comportementale basée sur la collecte massive de traces d'usage. Ces traces proviennent de différents dispositifs de tracking incluant les adresses IP (IP tracking), les désormais connus cookies ou les empreintes (p. ex. browser fingerprinting et canvas fingerprinting). Si la collecte s'est au départ limitée au poste de travail (essentiellement au travers du navigateur), elle a pu par la suite s'étendre aux smartphones et objets connectés. En a découlé le marketing des traces et l'économie de l'attention auxquels les digital natives ont été précocement confrontés. Diverses contres-mesures ont été progressivement déployées par les utilisateurs (paramétrage, extensions, p. ex. bloqueurs de publicités), par des services d'anonymisation (p. ex. VPN et proxy), par les éditeurs eux-mêmes ou par le régulateur (p. ex. RGPD). Ce papier exploratoire propose, d'une part, une présentation de la structuration du secteur de la publicité en ligne suivie par un état de l'art sur les outils de tracking qui y sont déployés, d'autre part, un inventaire et une analyse des contre-mesures déployées ainsi que de leur efficacité. Nous montrons en particulier l'évolution rapide des techniques utilisées et l'hétérogénéité de la couverture offerte par des dispositifs protecteurs a priori équivalents.

Mots-clés : marketing des traces, économie de l'attention, adtech, publicité programmatique, publicité comportementale, privacy, tracking, big data.

1. Introduction

Mesguish et Thomas (2013) distinguent quatre âges du web. Le premier, s'étendant de 1994 à 1996, est baptisé « *Web des pionniers* ». Cette expression désigne le développement d'un Web encore réduit en taille alimenté par des pionniers technophiles. De 1996 à 2004, le « *Web des documents* » s'accompagne d'une explosion du nombre de sites permise par la facilité des nouveaux outils d'édition de contenu et alimentée par les débuts du commerce électronique. La recherche d'information passe par les annuaires ou par les moteurs de recherche. Cette période voit la naissance de l'entreprise Google. Le « *Web social* », parfois appelé Web 2.0, s'étend de 2004 à 2010. Il voit une implication plus importante des utilisateurs dans la création et l'enrichissement des contenus. Enfin, le « *Web temps réel* » se développe dès 2010 avec la part croissante des réseaux sociaux (audience) ainsi que le développement des *smartphones* et des tablettes. Les applications mobiles se développent au détriment du Web classique (documents, hyperliens, etc.). Cette évolution s'est accompagnée d'une mutation de la publicité en ligne sous des formes de plus en plus ciblées (Peyrat, 2009), jusqu'à la publicité comportementale cherchant à coller au plus près des centres d'intérêt immédiats des consommateurs tels que révélés par leur historique de navigation. Cette personnalisation avancée suppose un travail permanent de *tracking* (p. ex. *cookies*) et d'analyse de données (profilage) par les régies publicitaires (p. ex. Google et Facebook). Ce profilage des utilisateurs couplé à la connexion permanente (via le *smartphone*) conduisent à une nouvelle forme de capitalisme basé sur l'économie de l'attention. Le concept d'économie de l'attention a fait l'objet d'un effort de théorisation de la part d'Emmanuel Kessous (Kessous, 2011 ; Kessous, 2012). Ce dernier décrit la transition d'un marketing de segmentation vers un marketing des traces renforçant l'emprise des offreurs sur les consommateurs en l'absence d'un contrôle fort des données à caractère personnel[1] par les individus. Dans un monde où le coût de l'accès à l'information tend vers 0, l'objet rare n'est plus l'information mais bien l'attention. La généralisation des activités d'extraction de traces d'usage conduit à la mise en place d'un capitalisme de surveillance (Zuboff, 2019) couvrant à la fois les mondes virtuels (p. ex. moteurs de recherche) et réels (p. ex. objets connectés).

Le secteur de la publicité en ligne a donc sensiblement évolué depuis ses débuts seconde moitié des années quatre-vingt-dix. Il a en particulier bénéficié des principales tendances technologiques liées à la transformation numérique, *cloud computing*, *big data* et *machine learning* en tête. A titre d'exemple, l'entreprise française Criteo possédait en 2015 plus de 10.000 serveurs répartis dans 6 centres de données permettant de traiter jusqu'à 800.000 requêtes HTTP par seconde (Clapaud, 2015). Il en a résulté une réorganisation progressive du secteur faite de concentration (p. ex. Google) mais aussi de spécialisation de certains acteurs plus

[1] Ces contributions ont été écrites avant la mise en œuvre par l'Union européenne du Règlement Général de Protection des Données (RGPD).

petits. Sur le plan du *tracking*, de nouvelles techniques apparaissent (p. ex. *device fingerprinting*) tandis que d'autres deviennent obsolètes compte tenu de l'apparition de nouvelles techniques ou de la diffusion de contre-mesures efficaces (p. ex. blocage par défaut du *canvas fingerprinting*). Face à cette débauche de mécanismes de pistage numérique et à l'omniprésence de la publicité, le secteur a cependant dû faire face à des réactions issues des consommateurs (p. ex. bloqueurs de publicités), des associations militantes (cf. Framablog, 2017) ou du législateurs (p. ex. réglementation pour la protection des données personnelles). Il existe donc un besoin pour un état des pratiques qui soit à jour en matière de publicité en ligne et d'outils de *tracking* prenant en compte leur efficacité au regard de la diffusion de contre-mesures technologiques (p. ex. bloqueurs de publicités) ou légales (p. ex. RGPD).

Ce papier exploratoire est décomposé en quatre sections. Dans une première section, nous proposons de dresser un panorama des pratiques avancées de publicité en ligne (publicité contextuelle, publicité comportementale, *retargeting*, publicité programmatique…). Elle sera suivie d'une section dédiée aux techniques de *tracking* que ces pratiques nécessitent. Dans une troisième section, nous dressons un inventaire des contre-mesures disponibles. Dans une quatrième section, et avant de conclure par les limitations et les perspectives de cette recherche préliminaire, nous discuterons la diffusion de ces contre-mesures et de leur efficacité.

2. Essor de la publicité programmatique

Le marketing en ligne s'appuie sur diverses techniques maintenant éprouvées : courriels commerciaux, réseaux sociaux numériques, référencement de sites internet... Parmi celles-ci, la publicité en ligne recourt principalement à la diffusion de bannières (*display*), dont les formats sont standardisés, et de liens sponsorisés (*search*) au sein des moteurs de recherche (Allary et al., 2018). Les transactions relatives aux bannières se sont pendant plusieurs années réalisées de gré à gré, conduisant surtout à la valorisation des espaces publicitaires présents dans les pages principales des sites web, dès lors entraînant de nombreux invendus parmi les espaces présents sur les pages secondaires (longue traîne). La valorisation de cet inventaire s'est dès lors ouvert aux réseaux publicitaires (*ad networks*, *affiliate networks* ; p. ex. Tradedoubler) offrant une rémunération moindre mais permettant d'améliorer substantiellement le taux de remplissage des espaces.

La publicité en ligne s'est progressivement sophistiquée avec la publicité ciblée. Peyrat (2009) en distingue trois variantes. La publicité personnalisée dites classique est adaptée « *en fonction des caractéristiques connues de l'internaute* » telles que son âge, son sexe ou sa localisation. Ces données sont fournies volontairement par l'internaute, par exemple lors de l'inscription sur un service. La publicité contextuelle est déterminée « *en fonction du contenu immédiat fourni à l'internaute* ». L'annonce affichée est donc adaptée au contenu de la page web sur laquelle elle est affichée. Le ciblage peut éventuellement être affiné grâce à la géolocalisation de l'internaute ou par la recherche d'information (requête) qui a

conduit à la page par le bais d'un moteur de recherche. La publicité comportementale est choisie « *en observant le comportement de l'internaute à travers le temps* ». En pratique, un profil individuel va être dressé sur base des d'actions (historique de visites de sites web, des mots-clefs rentrés dans les moteurs recherches…), permettant une adaptation des publicités proposées. Parmi les techniques éprouvées et diffusées, citons en particulier le *retargeting* (Allary et al., 2018 ; Lambrecht et al., 2013). Ce dernier permet l'affichage, sur des sites externes, d'une publicité liée à un produit proposé sur le site de l'annonceur et pour lequel l'internaute a, lors d'une visite sur le site, marqué un intérêt (visualisation d'une page, recherche par mot-clef, inclusion dans une liste d'envies ou un panier d'achats…). L'objectif est dès lors de raccompagner le prospect dans l'entonnoir de conversion (*funnel*) jusqu'à la concrétisation d'une action (p. ex. prise de contact ou ventes).

Plusieurs régies se sont spécialisées sur ces différentes techniques plus avancées. D'une part, Google a investi dès 2000 dans son service de publicité Google Adwords (rebaptisé Google Ads en 2018) permettant un affichage de publicités textuelles (liens sponsorisés[2]) adaptées aux mots-clefs soumis au moteur de recherche ainsi que, au travers de la régie Google Adsense, un affichage de publicités textuelles adaptées en fonction du contenu de la page web contenant l'espace publicitaire, des mots-clefs associés à la publicité (achetés aux enchères et facturés au CPC[3]), de la géolocalisation de l'internaute, de sa langue et de la plage horaire (Allary et al., 2018). D'autre part, la société française Criteo s'est différenciée par son service de *retargeting*[4] permettant la personnalisation des annonces en fonction des pages consultées (*retargeting* statique) ou d'un profil individuel dressé sur base de données comportementales exploitées par des algorithmes de *machine learning* (*retargeting* dynamique). Google a par la suite ajouté un service équivalent de remarketing dynamique à sa régie Google Ads[5].

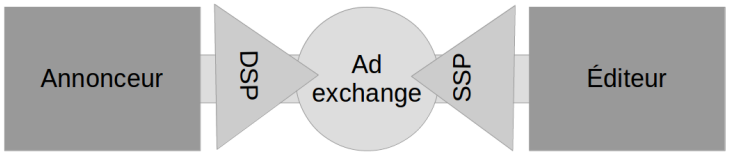


Figure 1. Écosystème de la publicité programmatique (Allary et al., 2013)

[2] Ce type de produit publicitaire est classé dans le *Search Engine Advertising* (SEA), distinct du référencement naturel, soit le *Search Engine Optimization* (SEO), les deux étant regroupés dans le *Search Engine Marketing* (SEM).

[3] CPC = *Cost per Clic* ou Coût par Clic.

[4] Cf. https://www.criteo.com/fr/quest-ce-que-le-retargeting-votre-guide-complet/ pour plus de détails.

[5] Cf. https://support.google.com/google-ads/answer/3124536 pour plus de détails.

Les pionniers comme Google ou Criteo ont ouvert la voie à une automatisation accrue de la publicité en ligne et ont conduit au développement de la publicité programmatique (Allary et al., 2018 ; cf. Figure 1). Cette dernière transforme la manière d'envisager une transaction commerciale entre un acheteur et un vendeur de publicité, au travers d'une place de marché (*ad exchange*) et grâce à la mise aux enchère en temps réel (RTB : *Real Time Bidding*) des espaces publicitaires disponibles (Renaud, 2017).

Tableau 1. Concentration et spécialisation des acteurs de la publicité programmatique (basé sur Allary et al., 2018 & Weide, 2018).

<table>
<tr><th>DSP</th><th>Ad exchange</th><th>SSP</th><th>Éditeur</th></tr>
<tr><td colspan="2">Google Adwords</td><td colspan="2">Google (search)</td></tr>
<tr><td colspan="2">Google Adwords</td><td>Google Adsense</td><td>Partenaires Adsense (display)</td></tr>
<tr><td>Search Ads 360 (ex-DoubleClick Search)</td><td>Google, Ads, Microsoft Advertising, Baidu...</td><td colspan="2">Bing, Baidu, Google, Yahoo (search)</td></tr>
<tr><td>Google Adwords, réseaux tiers</td><td>DoubleClick Ad Exchange (Google)</td><td colspan="2">Google Display Network, Adsense, Youtube, portails, sites d'actualités….</td></tr>
<tr><td colspan="3">Facebook Ads</td><td>Facebook</td></tr>
<tr><td>AdRoll, AppNexus, Criteo, MediaMath...</td><td>Criteo, DoubleClick (Google), Rubicon Project...</td><td>AppNexus, OpenX, PubMatic, Rubicon Project...</td><td>Éditeurs (portails, sites d'actualités…)</td></tr>
</table>

La publicité programmatique modifie profondément la chaîne de valeur de la publicité en ligne et voit l'émergence de fonctions spécialisées (Allary et al., 2018). Premièrement, l'annonceur consolide ses données clients au sein d'un DMP (*Data Management Platform*), notamment alimenté par ses outils CRM (*Customer Relationship Management*) et ses outils *analytics* permettant le suivi de l'activité sur les sites web de l'entreprise, éventuellement complété par les données fournies par des partenaires ou des courtiers en données (cf. Allary et al., 2018, et Framablog, 2017, pour plus de détails). Il peut ensuite émettre des ordres d'achat d'espace publicitaire sur un DSP (*Demand Side Platform*). A l'extrémité de la chaîne, les éditeurs de sites web gèrent un inventaire d'espaces publicitaires disponibles et diffusent des demandes d'offres (*bid requests*) sur un SSP (*Supplier-Side Platform*). Au centre, une plate-forme d'échange publicitaire (*ad exchange*) organise la rencontre entre les ordres d'achat (offre) et les demandes d'offres (demande) au travers d'un mécanisme d'enchères en temps réel (RTB). L'annonceur le plus généreux remporte l'enchère et son annonce peut dès lors être affichée sur le site de l'éditeur. Cette opération, dont la durée totale est inférieure à la seconde, suppose l'évaluation de la valeur commerciale de l'internaute face à l'espace publicitaire mis aux enchères (Allary et al., 2018 ; Framablog, 2017). Dans le cas idéal, les acteurs

spécialisés au sein de cette chaîne sont capables d'échanger des données (interopérabilité) et de mettre en commun des données relatives aux profils individuels, ce qui suppose un fastidieux travail de réconciliation de *cookies* (*cookie syncing*) et de création d'identifiants uniques (UUID) au sein notamment des DMP. Face à ces écosystèmes ouverts se positionnent les écosystèmes (partiellement) fermés de Google et Facebook (cf. Tableau 1).

3. Inventaire des techniques de *tracking*

La publicité en ligne s'appuie sur divers mécanismes de collecte de données personnelles et de suivi de l'activité des internautes au fil de leur navigation. Ce suivi passe par l'utilisation d'outils de *tracking*, une pratique qualifiée par ses détracteurs de « *pistage numérique* » (p. ex. Framablog, 2017). Un premier inventaire des mécanismes de *tracking* a été réalisé récemment par Ishtiaq et al. (2017).

Le *tracking* des adresses IP, aussi appelée IP *tracking*, permet le suivi de la navigation d'un internaute sur base de l'adresse IP reçue par chaque terminal de consultation connecté par Internet (Debize et al., 2016). L'adresse IP peut être fixe mais est plus généralement dynamique (p. ex. changement d'adresse lors du redémarrage d'une box internet domestique). L'IP *tracking* permet donc le suivi de la navigation sur une période de temps limitée. Cette méthode de *tracking* fonctionne par contre quelque soit le terminal (ordinateur personnel, téléphone…) et le logiciel d'accès à Internet (navigateur, application mobile...). L'adresse IP permet aussi la géolocalisation du terminal à l'échelle du pays (avec une fiabilité proche de 100%) ou de la ville (avec une fiabilité ne dépassant pas 90%) (Koch et al., 2013). En particulier, les adresses IP sont distribuées par l'ICANN par lots, l'appartenance à un lot permet de connaître l'organisation ou le pays correspondant à l'adresse.

La technologie centrale du *tracking* sur le Web est le *cookie* HTTP. Le *cookie* est un ensemble de données renvoyé par un serveur web au navigateur web et que ce dernier stocke ensuite localement[6]. Seul le serveur ayant créé le *cookie* HTTP peut ensuite en relire le contenu. De plus, les *cookies* ont une durée de vie limitée. Par ailleurs, ils peuvent être refusés (au cas par cas ou de manière systématique) ou supprimés à l'initiative de l'utilisateur (via les paramètres de sécurité du navigateur). S'il peut être utilisé pour gérer la connexion ou la personnalisation sur un site web, le *cookie* permet aussi le *tracking* à des fins publicitaires, soit qu'il contienne un identifiant permettant l'identification de l'utilisateur (et donc le rapprochement avec des données personnelles conservées en base de données par la régie) soit qu'il contienne des données relatives à son historique de navigation.

Le caractère potentiellement éphémère des *cookies* a conduit au développement de techniques pour en assurer la persistance. On parle alors de *cookie respawning*, d'*evercookie* voire de *cookie* zombie. Le principe consiste à récréer un *cookie* HTTP

[6] Cf. https://developer.mozilla.org/fr/docs/Web/HTTP/Cookies pour plus d'informations sur le fonctionnement technique des *cookies* HTTP.

après sa suppression en s'appuyant sur un autre dispositif de stockage, soit un *cookie* Flash (en réalité, un objet local partagé ou LSO), soit un mécanisme de stockage persistant dans le navigateur tel qu'IndexedDB (Acar et al., 2014). Le Flash n'étant plus utilisé que par moins de 3 % des sites web[7], le premier dispositif peut être considéré comme caduc.

Le *tracking* par *cookies* a été complété par diverses méthodes de calcul d'empreintes (*fingerprinting*), utilisables avec les navigateurs web (Acar et al., 2014), mais aussi avec les *smartphones*. S'agissant des navigateurs web, la technique consiste à exploiter l'extrême variété des configurations des navigateurs (*user agent* mais aussi liste des polices ou des extensions installées) et, plus largement, des postes de travail (système d'exploitation, modèle de carte graphique, version de pilote de carte graphique...). Le *browser fingerprinting* permet ainsi de calculer l'empreinte d'un navigateur sur base des spécificités précitées[8] tandis que le *canvas fingerprinting* exploite les différences (minimes) de rendu graphique. Plus précisément, le *canvas fingerprinting* consiste à transformer en image *lossless*, avec l'API canvas du navigateur web, une chaîne de caractères constituant un pangramme parfait (de manière à maximiser la diversité de rendu), puis à récupérer cette image avec la méthode Javascript `ToDataURL`, et enfin à transformer l'image en chaîne de caractères en utilisant le codage *base64*. Selon Acar et al. (2014), environ 5 % des sites classés dans le Top 100000 Alexa utilisaient le *canvas fingerprinting*, contre 2 % environ pour les sites issus du Top 1000 Alexa. Parmi les utilisateurs connus citons la société AddThis, dont les *widgets* sont largement diffusés et permettent une excellente couverture, soit 97,2 % selon Acar et al. (2014), de la population étasunienne. Firefox met en œuvre un blocage par défaut du *canvas fingerprinting* depuis Firefox 58 (publié le 23 janvier 2018).

En pratique, les méthodes de *fingerprinting* ont également été utilisées avec les téléphones (*device fingerprinting*). Elles s'appuient par exemple sur l'exploitation des données issues du suivi du rythme de décharge de la batterie (*battery fingerprinting*) ou des capteurs de mouvements (Chen et al., 2017 ; Das et al., 2018). Par ailleurs, les terminaux mobiles ont fait l'objet d'une nouvelle méthode de suivi : le *tracking* par ID (Allary et al., 2018 ; Reichgut, 2016). Ainsi, chaque terminal iOS (IDFA : *Identifier For Advertising*), Android (GAID : *Google Adversting ID*) ou Windows (WAID : *Windows Advertising ID*) possède un identifiant unique et non permanent, donc différent d'un numéro de téléphone ou d'un numéro de série, permettant le suivi du terminal (Al-Kabra et al., 2019*)*.

La collecte de données à caractère personnel peut aboutir à l'identification d'un individu au cours de sa navigation, soit de manière directe (authentification, ID…), soit de manière indirecte par croisement d'informations. Par exemple, à l'extrême, la géolocalisation d'une adresse couplée à une empreinte de navigateur peut conduire à l'identification d'un individu particulier si sa demeure est localisée dans une zone

[7] Cf. https://w3techs.com/technologies/details/cp-flash pour un suivi des statistiques d'utilisation.

[8] Cette technique peut notamment être testée avec le site Panopticlick développé par l'*Electronic Frontier Foundation* (EFF).

faiblement peuplée et qu'il utilise une configuration atypique sur son terminal de connexion. Les acteurs actifs dans la publicité en ligne, et en premier lieu les régies publicitaires, recourent par ailleurs à la synchronisation de *cookies* (*cookie syncing*) de manière à regrouper les informations collectées par différents serveurs (Acar et al., 2014 ; Papadopoulos et al., 2019). Cette activité est en particulier essentielle dans le contexte de la publicité programmatique (Allary et al., 2018).

4. Inventaire des contre-mesures

4.1. Configuration du navigateur

Par souci de simplification, cette section portera sur le navigateur Firefox. Ce dernier en est actuellement à la version 73.0 (publiée le 11 février 2020). Premièrement, la configuration du navigateur permet de limiter l'utilisation des *cookies* par les sites consultés, soit que l'utilisateur les refuse au fur et à mesure, soit que l'utilisateur en bloque certains de manière systématique, soit qu'il les supprime périodiquement. Au sein de Firefox, ces opérations peuvent être configurées dans l'onglet « Vie privée et sécurité ». Deuxièmement, la navigateurs offrent généralement une fonctionnalité de navigation privée. Cette dernière permet une navigation sans enregistrement des cookies et de l'historique de navigation au-delà de la session courante[9]. Troisièmement, certains navigateurs offrent des fonctionnalités avancées de blocage de *trackers*. C'est notamment le cas de Firefox avec la via la fonctionnalité *Enhanced Tracking Protection* (ETP) accessible depuis la barre d'adresse[10].

4.2. Installation d'extensions

Les navigateurs modernes permettent généralement l'installation d'extensions (*plugins*). Parmi les extensions populaires, citons les bloqueurs de publicités (p. ex. AdBlock Plus ou uBlock Origin). Les filtres mis en œuvre peuvent cependant dépendre du bloqueur utilisé. Édité par la société eyeo GmbH, AdBlock Plus filtre ainsi par défaut les serveurs publicitaires mis sur liste noire par la communauté EasyList mais laisse par contre passer des « *publicités acceptables* » c'est-à-dire conformes aux critères du Comité Publicité Acceptable d'où le service AdBlock Plus tire ses revenus… Parmi les « clients » de ce comité citons la société Criteo. Cette dernière ne manque d'ailleurs pas de mentionner (discrètement) sa porosité aux bloqueurs sur son site commercial (« *recover ad-blocked impressions with our ability to serve Acceptable Ads* »). La protection offerte par les bloqueurs de publicités varie donc d'une solution à l'autre.

Au côté des bloqueurs de publicités, d'autres extensions spécialisées sont proposées. Citons en particulier Ghostery. Ghostery s'appuie sur une base de

[9] Cf. https://support.mozilla.org/fr/kb/navigation-privee-naviguer-avec-firefox-sans-enregistrer-historique pour plus de détails.

[10] Cf. https://blog.mozilla.org/blog/2019/06/04/firefox-now-available-with-enhanced-tracking-protection-by-default/ pour plus de détails.

données de *trackers* (plus de 4500) classés par catégories (publicité, *analytics*, réseaux sociaux...) pour permettre, sur la plupart des navigateurs web du marché, le filtrage des *trackers* (ou de catégories de *trackers*) sélectionnés dans les paramètres de configuration de l'outil (par exemple, les boutons sociaux ne sont pas supprimés par défaut).

4.3. Protection par la législation

La protection des données à caractère personnel présente des approches distinctes en fonction des pays et des cultures. Trois pôles majeurs tendent ainsi à se dégager : les États-Unis, la Chine et l'Union européenne (Demiaux, 2018). Le modèle étasunien de régulation des données personnelles est davantage centré sur la primauté de la liberté individuelle, voire associe la *privacy* à un comportement de dissimulation et à une source d'inefficience (Rochelandet, 2010). La Chine permet pour sa part la collecte massive au profit tant de l'état[11] que des entreprises. Elle met d'ailleurs progressivement en place un régime de sanctions aux « mauvais » citoyens sur base d'un système de crédit social au sein duquel chaque citoyen chinois est associé à un score de réputation (Raphaël et al., 2019). Le modèle européen a divergé du modèle étasunien à partir des années soixante-dix en érigeant la protection des données à caractère personnel au rang de liberté fondamentale. Cette conception a conduit à la mise en application à partir du 25 mai 2018 du Règlement sur la Protection des Données Personnelles (RGPD). Le modèle européen prend en compte l'asymétrie de pouvoir entre les grands organismes et les citoyens, et veille au consentement éclairé des citoyens confrontés à la collecte de données personnelles.

Le RGPD[12] repose notamment sur des principes de consentement éclairé et de proportionnalité des données collectées au regard des finalités du traitement tels que communiquées à l'utilisateur. La notion de données à caractère personnel est large puisqu'elle inclut des données directement nominatives (telles que le nom et le prénom) et des données indirectement nominatives (Banck, 2018). Sont donc notamment couverts par le règlement les identifiants, les données de localisation, les adresses IP ou les *cookies* relatifs à une personne physique identifiable directement ou indirectement. Cette définition volontairement très large réduit sensiblement la marge de manœuvre des entreprises, obligées de demander à l'utilisateur une autorisation explicite et préalable à toute collecte de données à caractère personnel, désormais dans l'incapacité d'agir dans l'ombre sans risquer un constat de violation du règlement suivi d'une amende pouvant aller jusqu'à 10 millions d'euros ou 2 % du chiffre d'affaires annuel mondial de l'exercice précédent.

---

11 Nous ne développerons pas dans cet article la question de la collecte de données par les états et, en particulier, par les États-Unis. Nous renvoyons donc au chapitre 17 « *Cybersécurité : dimension géostratégique et politique* » de Debize et al. (2016) qui y consacrent un important développement.

12 Nous renvoyons à Banck (2018) pour une présentation complète mais synthétique du RGPD.

4.4. Anonymisation de la connexion

L'anonymisation de la connexion peut être mise en œuvre avec un niveau croissant d'efficacité par l'utilisation d'un *proxy*, d'un VPN ou d'un client Tor. Le *proxy* permet de masquer l'adresse IP du client car il expose sa propre adresse IP (Savchenko et al., 2015). Les VPN apportent en plus un chiffrement de la communication. Leur utilisation suppose de s'inscrire sur un serveur VPN à la fiabilité avérée, ce qui implique généralement le paiement d'un abonnement mensuel (p. ex. Ghostery Midnight). Quant à Tor, il repose sur une solution décentralisée et chiffrée s'appuyant sur un réseau de nœuds *proxy* (Savchenko et al., 2015). En outre, le navigateur Tor inclut différents mécanismes de lutte contre les *evercookies*, le *canvas fingerprinting* et le *cookie syncing* (Acar et al., 2014).

5. Discussion

L'utilisation de contre-mesures efficaces par les internautes suppose une conscience minimale des mécanismes de *tracking*. Ils se révèlent malheureusement sous informés. Si les internautes ont connaissance de l'existence de la collecte de données, la nature de cette dernière leur est souvent inconnue (Morey et al., 2018 ; cf. Tableau 2). Ainsi, les trois quarts des internautes ignorent la collecte de leur localisation alors que cette dernière peut être obtenue au travers des informations GPS (*smartphone*) ou de l'adresse IP du terminal. Dans le même ordre d'idée, selon une étude Connected Life 2017, « seuls » 29 % des Belges, 34 % des Français et 30 % des Européens utiliseraient un *adblocker*, contre 18 % des internautes dans le monde. Suire (2016) laisse par ailleurs entendre que l'installation d'un bloqueur chez les étudiants découle davantage d'un sentiment d'agacement face aux intrusions publicitaires que d'un rejet des pratiques de collecte massive de données à caractère personnel.

Tableau 2. Prise de conscience de la collecte de données (Morey et al., 2018).

Types de données	**Pourcentage d'individus conscients de partager ce type de données**
Liste d'amis sur les réseaux sociaux	27 %
Localisation	25 %
Recherche sur le web	23 %
Historique de communication (p. ex. archive de *chat*)	18 %
Adresses IP	17 %
Historique de navigation	14 %

Tableau 3. Évaluation de l'efficacité des contre-mesures.

	Firefox (configuration)	**Firefox (nav. priv.)**	**Bloqueur de publicité**	**Tor (client)**	**RGPD**
Portée	Générale	Générale	Publicité②	Générale	Juridique
*Cookie*①	✔	✔	✔	✔	✔
Evercookie	✘	?	✔	✔	✔
Browser fingerprinting	±③	±③	✔	✔	✔
Canvas fingerprinting	✔③	✔③	✔	✔④	✔
Adresse IP	✘	✘	✔	✔	✔
Historique de recherche	±	±	na	✔	✔
Réaction de l'éditeur	Aucune mais inconfort….	Détection et blocage⑤	Détection et blocage⑤	Détection et blocage⑤	Application partielle⑥

① Les *cookies* peuvent être facilement refusés et effacés, de manière manuelle ou automatique, à l'aide d'un navigateur web. La navigation privée permet de systématiser l'effacement des *cookies* à la fermeture de l'onglet de navigation.

② Les bloqueurs de publicité permet de bloquer l'affichage de la publicité mais aussi la collecte de données par le *tracker* (*tag*) en interdisant l'exécution du script Javascript correspondant. Par contre, il ne bloque pas d'autres types de *trackers* (p. ex. Google Analytics). Pour ces derniers, des extensions spécialisées doivent être installées au cas par cas ; Firefox, depuis la version 67.0.1, permet par ailleurs de configurer le blocage de *trackers* via la fonctionnalité *Enhanced Tracking Protection* (ETP)[13].

③ Le *canvas fingerprinting* est bloqué par Firefox depuis la version 58. L'énumération d'extensions (*plugins*) y est par ailleurs limitée[14].

④ Le client Tor inclut des contre-mesures permettant de lutter efficacement contre le *canvas fingerprinting* et le *cookie syncing* (Acar et al., 2014). De plus, Tor permet l'anonymisation de la connexion.

⑤ La détection d'une contre-mesure permet à l'éditeur de site web d'éventuellement bloquer l'affichage du contenu. La détection est notamment possible pour les bloqueurs de publicités, la navigation privée ou l'utilisation de Tor.

⑥ L'application du RGPD incombe uniquement aux organismes établis en Europe ainsi qu'aux organismes établis hors Union européenne traitant les données de citoyens européens. De plus, l'efficacité réelle dépend du caractère réellement éclairé du consentement de l'utilisateur, de l'activité de détection des infractions, des plaintes déposées et de la capacité (réduite) des autorités de contrôle nationales (p. ex. CNIL en France).

13 Cf. https://blog.mozilla.org/blog/2019/06/04/firefox-now-available-with-enhanced-tracking-protection-by-default/ pour plus de détails.

14 Cf. https://bugzilla.mozilla.org/show_bug.cgi?id=757726 pour plus de détails.

Le Tableau 3 propose une synthèse de contre-mesures courantes et analyse leur efficacité au regard des techniques de *tracking* et des contre-mesures potentiellement mises en place par les éditeurs de sites web. En pratique, le navigateur Firefox permet la mise en place, à la configuration, d'un large éventail de dispositifs pour limiter le pistage (effacement des *cookies*, détection des calculs d'empreintes, blocage de *trackers*, envoi d'entêtes HTTP « *Do Not Track* »...). Les mêmes fonctionnalités tendent à se retrouver sous Chrome (cf. Tableau 4). Cependant, intégré dans un écosystème plus large permettant à Google de collecter des données et de déployer ses services de publicités ciblées, Google Chrome organise une certaine perméabilité aidant l'entreprise à ainsi préserver son modèle d'affaires (p. ex. filtrage des « *publicités intrusives ou trompeuses* »).

Tableau 4. Comparaison de Chrome et Firefox (lutte contre le pistage).

Fonctionnalités de blocage		**Chrome**	**Firefox**
Popups		✓ (activé)	✓ (activé)
Cookies		✓ (configurable)	✓ (configurable)
Publicité	Natif	✓ (désactivé)	✗
	Extensions	✓ (Adblock Plus, Ghostery...)	✓ (Adblock Plus, Ghostery…)
Trackers	Natif	✗	✓ (désactivé)
	Extensions	✓ (« Désactivation de Google Analytics » + Ghostery, Privacy Badger...)	✓ (Ghostery, Privacy Badger...)
Do Not Track		✓ (désactivé)	✓ (désactivé)

La collecte non désirée de données peut s'apparenter à un problème de sécurité car elle viole la confidentialité des données à caractère personnel. Bien utilisée, les contre-mesures disponibles disposent d'une portée et d'une réelle efficacité, même si elles peuvent elles-mêmes s'exposer à des contre-mesures de la part notamment des éditeurs de sites web (p. ex. entrave à l'affichage d'une page en cas d'utilisation d'un bloqueur de publicités). Compte tenu de la puissance commerciale de Google, le choix des internautes en matière de navigateur web ne reflète cependant pas l'investissement des éditeurs en matière de *privacy* (cf. Tableau 5 ; statistiques : Statcounter) alors même que la dépendance de Google au modèle d'affaires publicitaire est régulièrement rappelée (cf. par exemple Nitot, 2016).

Tableau 5. Comparaison des navigateurs.

	Open source	**Diffusion**	**Mise à jour**	**Innovation**	**Publicité ciblée**	***Privacy***
Firefox	✓	5 %	⋆⋆⋆	⋆⋆⋆		⋆⋆⋆
Chromium	✓	< 0,1 %	⋆⋆	⋆⋆		⋆⋆
Chrome		60 %	⋆⋆⋆	⋆⋆⋆	✓	⋆
Safari		15 %	⋆⋆⋆	⋆⋆		⋆⋆
Internet Explorer		< 2 %	⋆	⋆		⋆⋆
Edge		< 5 %	⋆⋆⋆	⋆⋆		⋆(⋆)
Opera		2,5 %	⋆⋆⋆	⋆⋆⋆	✓	⋆

Le extensions pour les navigateurs sont diversifiées et diffèrent suivant différents critères : couverture, compatibilité et confort d'utilisation (cf. Tableau 6 pour quelques exemples). Deux points ressortent. D'une part, à côté de solutions partielles existent des solutions globales permettant de configurer le filtrage de différents types de *trackers* (p. ex. Ghostery). D'autre part, à l'instar des navigateurs web, l'effectivité du filtrage est dépendante du modèle d'affaires de l'éditeur et de ses liens avec le marché de la publicité ciblée (p. ex. Adblock Plus) !

Tableau 6. Efficacité des extensions.

Extension	**Automatique**	**Couverture**	**Configuration**	**Compatibilité**	**Désagrément(s) connu(s)**	**Intérêt**
Adblock Plus	✓	*Trackers* publicitaires	✓	Chrome, Firefox, Internet Explorer, Safari, Edge, Opera…	Acceptation de la « *publicité acceptable* », détection par les sites	⋆
Ublock Origin	✓	*Trackers* publicitaires	✓	Chrome, Safari, Firefox, Chromium	Refus d'inclusion dans Chrome Web Store	⋆⋆
Ghostery	✓	*Trackers* publicitaires, *trackers analytics…*	✓	Chrome, Firefox, Safari, Edge, Opera, Cliqz (Firefox)	Détection (épisodique) par les sites	⋆⋆⋆
Privacy Badger	✓	*Trackers* (dont publicitaires)	✓	Chrome, Firefox, Opera	Ralentissement de la navigation (?)	⋆⋆

6. Conclusion

Notre recherche exploratoire nous a permis de décrire les évolutions du secteur de la publicité et d'expliquer le besoin en techniques avancées de *tracking* permettant le suivi individualisé de la navigation mais aussi la création de profils et la mise en commun de données via la synchronisation de *cookies*. Sur base d'un inventaire complet des techniques de *tracking*, nous avons pu proposer une analyse de l'efficacité de ces mécanismes de *tracking* ainsi que des contre-mesures déployées pour limiter ou bloquer la collecte de données à caractère personnel. Nous avons notamment montré l'hétérogénéité de la protection offerte par des contre-mesures à première vue équivalentes (p. ex. bloqueurs de publicités).

Notre recherche souffre d'au moins quatre limitations. Premièrement, l'analyse des possibilités de configuration ou d'extension des navigateurs s'est focalisée sur le navigateur Firefox. Or, la Mozilla Foundation, qui le produit, se distingue par ses engagements en faveur de la protection de la vie privée[15]. Une analyse similaire devrait donc être réalisée pour les éditeurs d'autres navigateurs web (p. ex. Apple Safari et Microsoft Edge), en particulier ceux édités par des entreprises dont le modèle d'affaires dépend de la publicité ciblée (p. ex. Google Chrome et Opera). Deuxièmement, le poste de travail fait l'objet d'une collecte de données au départ du système d'exploitation voire aussi des applications installées. Les activités de télémétrie liées à l'amélioration de la qualité sont ainsi critiquées (p. ex. télémétrie sous Windows 10) tandis que certaines applications sont épinglées pour la revente de données à caractère personnel éventuellement anonymisées (cf. antivirus Avast[16]). Troisièmement, notre recherche est centrée sur le poste de travail et les navigateurs web qui y sont utilisés. Or, l'utilisation d'Internet s'est substantiellement déplacée vers les terminaux mobiles (*smartphones*, tablettes…). Ces derniers présentent des particularités techniques et sont soumis à une activité très importante de collecte de données via des *trackers* publicitaires intégrés aux *apps* (Binns et al., 2018) ainsi que, *last but not least*, *via* des logiciels pré-installés tels qu'Android ou Google Maps (Nitot, 2016). Quatrièmement, l'analyse des outils de *tracking* investigue peu les techniques récentes dédiées aux terminaux mobiles et aux objets connectés de type identifiant unique non permanent. Cette pratique, largement diffusée pour les terminaux mobiles, intéresse également dans le cadre du développement de la publicité programmatique sur les télévisions connectées et s'inscrit dans une réflexion plus large sur le remplacement des *cookies* tiers dont le filtrage par les navigateurs (p. ex. Firefox et Safari) est de plus en plus fréquent (Sluis, 2020).

Parmi les perspectives, outre le travail sur les limitations précitées, citons l'élaboration d'une méthodologie outillée permettant de mesurer l'exposition d'un individu à la collecte de données à caractère personnel compte tenu de ses usages de dispositifs connectés et des éventuelles contre-mesures mises en œuvre.

---

15 Cf. https://www.mozilla.org/fr/firefox/privacy/ pour plus de détails.

16 Cf. https://www.zdnet.fr/actualites/oui-avast-vend-des-donnees-personnelles-et-cela-se-savait-39898209.htm pour plus de détails.

Bibliographie

Acar G., Eubank C., Englehardt S., Juarez M., Narayanan A. & Diaz C. (2014), The web never forgets: Persistent tracking mechanisms in the wild. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, p. 674-689.

Al-Kabra R., Bodiga P. K., Dahlstrom N., Sinha R., Morrow J., Drake A., & Phan C. (2019). *Ascertaining network devices used with anonymous identifiers*. U.S. Patent Application No. 15/801,971.

Allary J. & Balusseau V. (2018). *La publicité à l'heure de la data. Adtech et programmatique expliqués par des experts*, Dunod.

Baudry B. & Laperdrix P. (2015). *Le fingerprinting : une nouvelle technique de traçage*, MISC, n°081, septembre 2015. En ligne : https://connect.ed-diamond.com/MISC/MISC-081/Le-fingerprinting-une-nouvelle-technique-de-tracage.

Banck A. (2018). *RGPD : la protection des données à caractère personnel*, Gualino.

Binns R., Lyngs U., Van Kleek M., Zhao J., Libert T. & Shadbolt N. (2018). Third party tracking in the mobile ecosystem. In *Proceedings of the 10th ACM Conference on Web Science*, p. 23-31.

Broussard G. (2019). *Internalisation programmatique en France : taux d'adoption, avantages, degrés et types de fonction d'achat intégré par rapport à l'Europe*, Interactive Advertising Bureau (IAB), avril 2019.

Chen J., Fang Y., He K. & Du R. (2017). Charge-Depleting of the Batteries Makes Smartphones Recognizable, *2017 IEEE 23rd International Conference on Parallel and Distributed Systems (ICPADS)*, Shenzhen, p. 33-40. DOI : 10.1109/ICPADS.2017.00016.

Clapaud A. (2015). *Criteo, une architecture Big Data unique au monde*, Le Journal du Net, 10 mars 2013. En ligne : https://www.journaldunet.com/solutions/cloud-computing/1151178-criteo-une-architecture-big-data-unique-au-monde/.

Das A., Acar G., Borisov N. & Pradeep, A. (2018). The Web's Sixth Sense: A Study of Scripts Accessing Smartphone Sensors. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, p. 1515-1532.

Debize T., Anzala-Yamajako A., Soullié A., Billois G., Kokos A., Wolfhugel C. & Bloch L. (2016). *Sécurité informatique: Pour les DSI, RSSI et administrateurs*, Eyrolles.

Demiaux V. (2018). *De la CNIL au RGPD : 40 ans de protection des données (interview)*, L'Histoire, 25 mai 2018. En ligne : https://www.lhistoire.fr/entretien/de-la-cnil-au-rgpd%C2%A0-40-ans-de-protection-des-donn%C3%A9es.

Framablog (2017). *Comment les entreprises surveillent notre quotidien*, Framablog, 25 octobre 2017. En ligne : https://framablog.org/2017/10/25/comment-les-entreprises-surveillent-notre-quotidien/.

Ishtiaq A., Abbasi S. H., Aleem M., & Islam M. A. I. (2017). *User tracking mechanisms and counter measures*. International Journal of Applied Mathematics Electronics and Computers, 5(2), p. 33-40.

Kessous E. (2011), L'économie de l'attention et le marketing des traces, *Actes du colloque Web social, communautés virtuelles et consommation*.

Kessous E. (2012). *L'attention au monde. Sociologie des données personnelles à l'ère numérique*, Armand Colin.

Koch R., Golling M. & Rodosek G. D. (2013). Advanced geolocation of IP addresses. In *International Conference on Communication and Network Security (ICCNS)*, p. 1-10.

Kosinski M., Stillwell D. & Graepel T. (2013). *Private traits and attributes are predictable from digital records of human behavior*, PNAS April 9, 110(15) p. 5802-5805. En ligne : https://www.pnas.org/content/110/15/5802.

Lambrecht A. & Tucker C. (2013). *When Does Retargeting Work? Information Specificity in Online Advertising*. Journal of Marketing Research, 50(5), p. 561–576.

Mesguish V. & Thomas A. (2013). *Net recherche 2013*. De Boeck. ISBN : 978-2-8041-8228-1.

Morey T., Forbath T. & Schoop A. (2018). *Données clients : concevoir un système transparent de confiance*, Harvard Business Review, Printemps 2018, p. 64-74.

Nitot, T. (2016). *surveillance://*, C&F éditions. ISBN : 978-2-915825-65-7.

Papadopoulos P., Kourtellis N. & Markatos E. (2019). Cookie synchronization: Everything you always wanted to know but were afraid to ask. In *The World Wide Web Conference*, p. 1432-1442.

Peyrat B. (2009). *La publicité ciblée en ligne*, CNIL. En ligne : https://www.cnil.fr/sites/default/files/typo/document/Publicite_Ciblee_rapport_VD.pdf.

Raphaël R. & Xi L. (2019). *Quand l'état organise la notation de ses citoyens. Bons et mauvais chinois*, Le Monde diplomatique, janvier 2019. En ligne : https://www.monde-diplomatique.fr/2019/01/RAPHAEL/59403.

Reichgut M. (2016), *Advertiser ID Tracking And What It Means For You*, Forbes, 16 mai 2016. En ligne : https://www.forbes.com/sites/onmarketing/2016/05/16/advertiser-id-tracking-and-what-it-means-for-you/#c8d03a118bf0.

Renaud J.-F. (2017). *Les achats programmatiques : comprendre les enjeux*, Gestion, 2017/2 (Vol. 42), p. 106-109. DOI : 10.3917/riges.422.0106. En ligne : https://www.cairn.info/revue-gestion-2017-2-page-106.htm.

Rochelandet F. (2010). *Économie des données personnelles et de la vie privée*, La Découverte, Paris.

Savchenko, I.I., Gatsenko, O.Y. (2015). *Analytical review of methods of providing internet anonymity*. Aut. Control Comp. Sci. 49, 696–700 (2015).

Sluis S. (2020). *Post-Cookie Apocalypse, IAB Unveils 'Project Rearc'*. AdExchenger, 11 février 2020. En ligne : https://www.adexchanger.com/ad-exchange-news/post-cookie-apocalypse-iab-unveils-project-rearc/.

Suire R. (2016). *GénérationY, GénérationZ, Génération A-nalphanète ? Portrait d'une cohorte d'étudiants en 2016*. M@rsouin, Université de Rennes.

Weide K. (2018), *Worldwide Digital Advertising Software Market Shares, 2017: Despite Intense M&A Activity, Still a Fragmented Market*, IDC, septembre 2018.

Zuboff S. (2019). *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs. ISBN : 978-1610395694.

Quelle Blockchain choisir ? Un outil d'aide à la décision pour guider le choix de technologie Blockchain

Nicolas Six, Nicolas Herbaut, Camille Salinesi

Centre de Recherche en Informatique – EA 1445
Université Paris 1 Panthéon-Sorbonne
nicolas.six,nicolas.herbaut,camille.salinesi@univ-paris1.fr

RÉSUMÉ. *Les entreprises qui souhaitent déployer des solutions basées sur la blockchain sont confrontées à une pléthore de technologies concurrentes ayant chacun un grand nombre de paramètres propres devant être ajustés par un expert. Dans cet article, nous proposons un algorithme d'aide à la décision permettant de mieux prendre en compte les exigences haut niveau et préférences pour les recommandations. Nous construisons une base de connaissance de solutions blockchain et exécutons un processus de décision multicritère automatisé donnant la solution la plus pertinente à partir des exigences et préférences. Nous validons notre approche sur un cas d'étude de gestion de chaîne logistique. Nous donnons des pistes pour une prise en compte plus flexible des exigences dans de futurs travaux.*

ABSTRACT. *Companies trying to build new solutions using blockchain are confronted with a plethora of available concurrent technologies that have many control knobs which require fine-tuning by experts. Exiting studies that build decision models for blockchain adoption or selection lack an automated way to use non-functional requirements to provide recommendations. In this paper, we build a knowledge base for blockchain solutions by analyzing whitepapers and studies, but also our benchmark results performed in a controlled environment. Then, we implement a Multi-Criterion Decision Analysis method to determine the most suitable blockchain solution from companies provided requirements and preferences. Finally, we illustrate our approach by running the decision process on a realistic supply-chain use case. This paper provides a rationale for blockchain deployment choices. While still limited in scope, we plan to include more blockchain alternative and more flexible requirements inputs in future work.*

MOTS-CLÉS : *Blockchain, Ingénierie des exigences, Aide à la décision multicritère*

KEYWORDS: *Blockchain, Requirements engineering, Multi-criteria decision analysis*

1. Introduction

La blockchain (ou chaîne de blocs) est un registre distribué maintenu à jour par un ensemble de nœuds. Les utilisateurs peuvent interagir avec les nœuds afin d'envoyer à la blockchain des transactions. Un registre blockchain prend la forme d'un ensemble de blocs contenant les transactions soumises par les utilisateurs, ainsi que des métadonnées sur lui-même ou le registre. Chaque bloc est relié au précédent par la valeur de hachage (en anglais, *hash*) de celui-ci. Dans ce sens, comme la modification d'un bloc altérerait cette valeur ainsi que toutes les autres valeurs de hachage des blocs suivants, il est théoriquement impossible d'altérer le contenu d'un bloc. Les nouveaux blocs sont formés par un sous-ensemble de nœuds chargés de regrouper les transactions dans un bloc et de le valider en mettant en œuvre différents processus cryptographiques en fonction de la blockchain utilisée, afin de garantir sa validité lors de son ajout à la blockchain.

La blockchain est apparue lors de la création de Bitcoin (Nakamoto, 2008), afin de permettre aux utilisateurs d'échanger la cryptomonnaie de même nom. Par la suite, de nombreuses blockchains ont pu voir le jour. Ethereum, la plus connue, est plébiscitée (Wood *et al*., 2014) pour sa capacité à déployer et à interagir avec des contrats intelligents (en anglais, smart contracts), logés dans la blockchain (Szabo, 1997). Les contrats intelligents pour la blockchain permettent non seulement d'exécuter des fonctions directement au sein de celle-ci, mais aussi de stocker des états. Ils bénéficient donc directement des propriétés particulières de la blockchain, qui sont intégrité, décentralisation, non-répudiation des transactions et transparence. Cela donne à la blockchain un statut de tiers de confiance artificiel, où il est possible de faire confiance au code et à la puissance du réseau contrairement aux tiers de confiance conventionnels qui assurent la validité des transactions au travers de leur statut, tels que les banques ou les gouvernements.

Ces caractéristiques ont attiré l'attention des industriels et universitaires, qui voient en la blockchain un moyen de révolutionner la manière d'échanger de la valeur entre individus ainsi que de garantir la véracité et l'intégrité des données stockées dans celle-ci. En effet, la blockchain serait "un support numérique natif pour la valeur, par lequel nous pourrions gérer, stocker et échanger de multiples biens [...] de pair à pair et de manière sécurisée" (Tapscott, 2016). De ce fait, on trouve dans la littérature de nombreux cas d'utilisations pertinents de la blockchain dans différents secteurs d'activités, tels que la gestion de chaîne logistique (Abeyratne, Monfared, 2016), la finance (Hyvärinen *et al*., 2017), le contrôle du réseau (Herbaut, Negru, 2017), l'identité décentralisée numérique (Takemiya, Vanieiev, 2018) ou encore la santé (Ekblaw *et al*., 2016).

Cependant, malgré son potentiel, la blockchain fait face à de nombreux freins à l'adoption. D'après une étude réalisée par l'entreprise PwC en 2018 [1], les entreprises se heurtent à des problèmes tels qu'une régulation incertaine sur le sujet, un manque

1. https://www.pwc.com/gx/en/issues/blockchain/blockchain-in-business.html

de confiance envers les autres acteurs lors de leur participation à un projet utilisant la blockchain, ainsi qu'à la difficile gestion de la propriété intellectuelle des données et biens qui y sont enregistrés. Ces problèmes sont résolus petit à petit grâce à la collaboration des acteurs de l'écosystème blockchain et des instances juridiques et gouvernementales compétentes. Néanmoins, les entreprises sont encore confrontées à un frein technologique, et ce pour plusieurs raisons. Elles peuvent rencontrer des difficultés à recruter des collaborateurs spécialisés en blockchain, la technologie étant encore jeune. Elles peuvent aussi avoir du mal à intégrer la blockchain à leurs systèmes d'information et processus métiers existants, car il n'existe pas encore de bonnes pratiques identifiées et éprouvées en entreprise par les architectes logiciels. Afin de pallier ce problème, des études ont été menées afin d'assister l'intégration de la blockchain dans des architectures logicielles. Dans ce sens, une étude propose une collection de modèles architecturaux contenant de la blockchain, ainsi que les différents cas où ces modèles sont applicables (Xu *et al*., 2018).

Mais le frein principal se situe sur la conception de la solution blockchain ainsi que son implémentation. À ce stade, les développeurs peuvent se poser plusieurs questions. Quelle blockchain utiliser dans un contexte donné, sachant qu'il existe de nombreuses technologies concurrentes avec, pour chacune, des propriétés et caractéristiques qui leur sont propres ? Peut-être est-ce finalement plus raisonnable d'utiliser une solution "éprouvée" au lieu d'une blockchain (base de données, microservices ...) ? Enfin, comment configurer les différents paramètres de la blockchain, qui ont un impact important sur la satisfaction des exigences (performances, résilience, sécurité ...) tels que l'algorithme de consensus ou l'intervalle inter-blocs, nécessitant souvent l'intervention d'experts dans le domaine pour aboutir à un résultat satisfaisant les exigences ?

De nombreuses études ont été menées pour répondre aux deux premières questions et ainsi facilitent le choix de la solution blockchain, notamment par le biais de modèles de décision à travers diverses questions (Wust, Gervais, 2018 ; Koens, Poll, 2018). Une autre étude (Belotti *et al*., 2019) présente un *vadémécum* contenant toutes les informations nécessaires à la compréhension de la blockchain d'un point de vue technique, ainsi qu'un modèle de décision pour la blockchain appliqué à plusieurs scénarios d'exemple. Les systèmes de recommandation proposés sont souvent constitués d'une série de questions abstraites, permettant de répondre à des problématiques telles que "ai-je besoin d'une blockchain ?" ou "quel type de blockchain adopter ?", mais pas de fournir des recommandations précises, ou rentrent plus dans le détail en considérant les choix entre de nombreux paramètres et propriétés blockchain. Les utilisateurs souhaitant obtenir une recommandation plus précise doivent donc se tourner vers ces derniers. Ce type d'étude est pertinent pour des personnes ayant de bonnes connaissances dans le domaine de la blockchain, mais il sera difficile pour des personnes non initiées de répondre de manière précise aux questions du modèle de décision. De plus, beaucoup de ces études se concentrent uniquement sur les exigences blockchain, alors que les utilisateurs ont des exigences liées à la qualité logicielle (performance, sécurité, fiabilité ...). Les liens qui relient les attributs blockchain aux qualités logicielles telles que définies en ingénierie, sont souvent peu explicites et il est difficile de quantifier l'impact d'un paramètre blockchain sur les qualités logicielles de la solution

finale. Enfin, lorsque le nombre d'attributs techniques considérés devient important, il est impossible de réaliser un choix les prenant tous en compte, la complexité de calcul lors de l'utilisation manuelle du modèle étant trop élevée.

Pour pallier ces limitations, nous introduisons dans cet article un processus de décision automatisé qui détermine l'alternative la plus intéressante pour un cas d'étude donné. Dans celui-ci, les préférences ou exigences des utilisateurs quant à la qualité logicielle de la solution à créer seront utilisées en entrée. Celles-ci seront comparées aux différentes caractéristiques des alternatives considérées par une méthode d'aide à la décision multicritère. Ces caractéristiques seront contenues sous la forme d'une base de connaissance et définies en utilisant la littérature existante (expériences, revues de littérature ...), les livres blancs des blockchains considérées ainsi que nos propres résultats d'expériences. Nous présentons également une application de notre processus de décision à un cas d'utilisation pertinent dans le domaine de la gestion de chaîne logistique. Cette partie sera l'occasion pour nous de valider les résultats du processus de décision, par le biais d'expériences manuelles confirmant les décisions prises par le processus.

La section 2 de cette étude est consacrée au processus de décision, la section 3 sur l'application du processus au cas d'étude sur la gestion de chaîne logistique. Nous présentons les travaux connexes à notre étude dans la section 4, puis nous enchaînons sur une discussion quant à nos résultats et notre approche dans la section 5. Enfin, nous concluons notre étude et introduisons nos travaux futurs dans la section 6.

2. Construction du processus de décision

Dans cette section, nous allons présenter les entrées ainsi que le fonctionnement du processus de décision permettant d'aider l'utilisateur à choisir le modèle de blockchain le plus approprié.

2.1. Entrées

La précision d'un algorithme d'aide à la décision multicritère dépend majoritairement des données saisies en entrée. Dans cette sous-section, nous présentons notre approche pour construire une base de connaissance fiable et adaptée, ainsi que notre méthode pour éliciter les poids qui seront appliqués à chacun des critères pour l'exécution du processus de décision.

2.1.1. Alternatives et attributs

Pour alimenter notre processus d'aide à la décision, nous avons construit une première version de base de connaissance contenant un ensemble d'alternatives de blockchains a_m et de leurs attributs respectifs c_n (Tableau 1). Nous avons choisi ce panel spécifique de blockchains, car (hors Bitcoin) elles sont considérées comme les blo-

ckchains les plus utilisées par les fournisseurs de service blockchain en entreprise [2]. Cependant, nous avons quand même choisi d'inclure la blockchain Bitcoin dans notre base de connaissance, car elle la plus connue du grand public, mais aussi la plus ancienne.

L'objectif de notre travail étant d'aider les entreprises à prendre des décisions sur la blockchain à utiliser sans avoir d'expertise particulière quant à la configuration de celle-ci, nous avons choisi un ensemble de critères qui peuvent être placés sous les différents macro-attributs proposés par la norme ISO 25010 [3], un standard définissant les différents macro-attributs à considérer afin de garantir la qualité d'un système ou d'un logiciel lors de son implémentation. Nous avons choisi les attributs qui nous semblent pertinents dans les considérations à avoir lors du choix d'une blockchain, mais aussi pour la possibilité à les retranscrire sous format numérique. Par conséquent, nos critères ne sont pas spécifiques à la technologie blockchain, mais relatifs à la qualité système. C'est notre processus de décision qui aura la charge de transcrire ces attributs de qualité système en attributs blockchain (tels que l'intervalle inter-blocs, l'algorithme de consensus, ou la taille des blocs). La figure 1 présente un diagramme indiquant les relations entre les attributs de qualité logicielle (critères choisis pour notre processus de décision) et les attributs spécifiques à la blockchain. Les valeurs saisies pour chacun

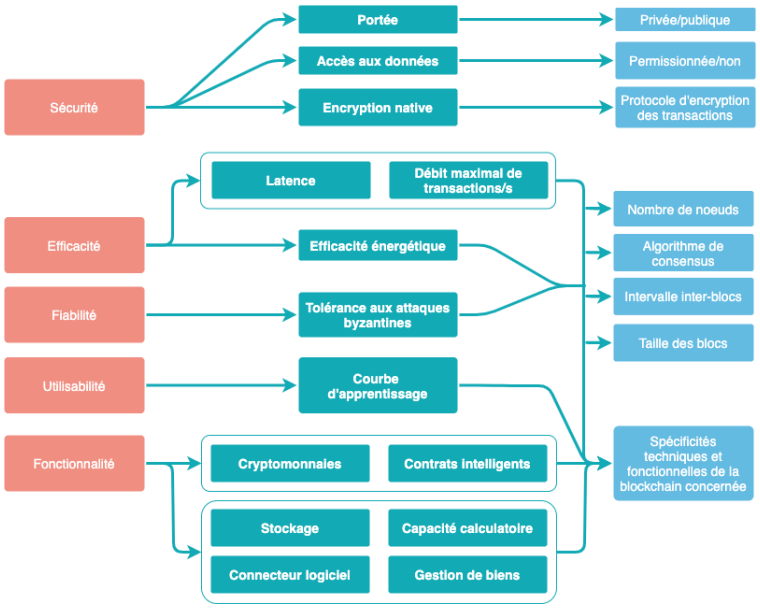


Figure 1. Attributs choisis (milieu) reliés aux qualités système (à gauche) et blockchain (à droite).

2. https://www.hfsresearch.com/pointsofview/whos-winning-the-battle-of-enterprise-blockchain-platforms
3. https://iso25000.com/index.php/en/iso-25000-standards/iso-25010

des attributs d'alternatives de notre base de connaissance proviennent de différentes sources : études (comme celle de (Belotti *et al*., 2019)), livres blancs (e.g. (Brown *et al*., 2016), (Nakamoto, 2008), (Wood *et al*., 2014)), documentations techniques et littérature scientifique (e.g. (Androulaki *et al*., 2018)).

Certaines de ces valeurs sont approximatives (marquées par le symbole $\mp$), car soumises à variation, de la topologie et la configuration du réseau blockchain ainsi que des caractéristiques techniques des nœuds le composant (CPU, mémoire vive ...). Leur valeur est donc construite à partir d'attributs connus, comme l'algorithme de consensus supporté (un algorithme tolérant les fautes byzantines comme l'algorithme PoW de Bitcoin aura un débit de transaction plus faible qu'un algorithme tolérant aux pannes, comme Raft utilisé par Hyperledger Fabric). Néanmoins, ces valeurs peuvent être fixées lorsque les paramètres blockchain sont connus. Notre processus de décision devant prendre en compte des actifs déjà présents dans l'entreprise (comme l'infrastructure technique ou les modèles de processus métiers), nous comptons sur la réalisation de tests de performance afin de pouvoir donner une valeur fixe aux attributs variables en fonction du contexte donné. Cette base de connaissance sera également variable dans le temps. Les valeurs des attributs des différentes blockchains choisies seront modifiées si nécessaire (mise à jour d'un des éléments d'une blockchain). Ces variations pouvant avoir un impact sur le choix de la meilleure alternative par notre processus de décision, il sera nécessaire d'évaluer l'ancienneté de la base de connaissance afin de déterminer si la recommandation est pertinente à un instant donné.

Tableau 1. Alternatives et attributs retenus.

Attributs/Alternatives	**Bitcoin**	**Ethereum**	**Ethereum**	**Hyperledger Fabric**	**Corda**
Algorithme de consensus	PoW [a]	PoW	PoA [b]	Raft	PBFT [c]
Ouvert publiquement	Oui	Oui	Non	Non	Non
Permissions	Non	Non	Non	Oui	Oui
Encryption native	Non	Non	Non	Oui	Oui
Débit (tx/s)	3,8	15	$\mp$100	$\mp$1000	$\mp$1000
Latence (s)	3600	180	$\mp$10	<1	<1
Efficient en énergie	Non	Non	Oui	Oui	Oui
Tolérant aux fautes byzantines	50,00%	50,00%	33,30%	0,00 %	33,30%
Contrats intelligents	Non	Oui	Oui	Oui	Oui
Cryptomonnaies	Oui	Oui	Oui	Non	Non
Element de stockage	Basique	Avancé	Avancé	Avancé	Avancé
Elément de calcul	Non	Avancé	Avancé	Avancé	Avancé
Elément gestionnaire de biens	Basique	Avancé	Avancé	Avancé	Avancé
Connecteur logiciel	Non	Avancé	Avancé	Avancé	Avancé
Courbe d'apprentissage	Faible	Moyenne	Moyenne	Très élevé	Très élevé

a. Proof-of-work (PoW), preuve de travail

b. Proof-of-Authority (PoA), preuve d'autorité

c. Practical Byzantine Fault Tolerance

2.1.2. Poids et conditions définis par l'utilisateur

Afin d'obtenir une préconisation de blockchain conforme aux attentes de l'utilisateur, le processus de décision automatisé doit prendre en compte les exigences

et préférences de celui-ci. Lorsque l'utilisateur est invité à saisir ses choix, il peut marquer un critère comme *Requis* ou *Indésirable*. Lors de la prise de décision, une alternative dont l'attribut ne respecterait pas l'une de ces deux exigences serait automatiquement disqualifiée des alternatives possibles, indépendamment de son score obtenu par l'exécution de l'algorithme d'aide à la décision multicritère.

L'utilisateur peut également indiquer ses préférences quant aux attributs, par le biais de variables littérales formant une échelle de Likert (Allen, Seaman, 2007) (Tableau 2). Le choix d'une de ces variables permet d'obtenir une valeur de préférence $p_n \in \ltimes$ pour chacun des critères c_n. Afin d'obtenir les poids de chacun des critères ω_n de telle façon à ce que la somme de ces poids soit égale à 1, il nous faut diviser chacune des préférences p_n pour un critère par la somme des préférences.

Tableau 2. Échelle de Likert associant labels et valeurs de préférence.

Variable linguistique	**Valeur de préférence** p_n
Extrêmement désirable	4
Tout à fait désirable	3
Désirable	2
Faiblement désirable	1
Indifférent	0

2.2. Logique interne

Tout d'abord, notre processus de décision effectue un premier filtrage des alternatives en fonction des exigences de l'utilisateur. Si un critère marqué comme *Requis* ou *Indésirable* n'est pas respecté par l'une des alternatives, elle est automatiquement éliminée, peu importe le score qu'elle aurait pu obtenir à l'aide de l'algorithme de décision qui suit. Pour un critère *Requis* qui n'est pas un booléen, l'utilisateur spécifie une valeur extremum. Par exemple, si un certain nombre de transactions par seconde est requis, les alternatives qui n'atteignent pas la valeur seuil seront disqualifiées.

Le processus de décision automatisé sur les alternatives restantes repose sur l'utilisation d'un algorithme d'aide à la décision multicritère appelé TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) (Lai *et al.*, 1994). L'algorithme TOPSIS est basé sur le fait que l'alternative a_m la plus pertinente pour un ensemble de choix donné doit être la plus proche possible de la solution idéale positive A^+ et la plus éloignée de la solution idéale négative A^-.

Le choix de cet algorithme a été guidé grâce à une étude présentant un état de l'art des études portant sur le choix d'une méthode d'aide à la décision multicritère (Kornyshova, Salinesi, 2007). Celle-ci propose un cadre de décision incluant différentes propriétés sur lesquelles porter notre attention lors du choix d'une méthode d'aide à la décision multicritère. Nous avons jugé que la méthode TOPSIS était adaptée à notre processus de décision, notamment car elle supporte l'analyse multicritère d'attributs nombreux et variés (ce qui est le cas lors de la comparaison de deux blockchains) tout en étant simples d'implémentation et précise dans la décision. Elle permet

également de prendre en compte des poids définis par un utilisateur, ce qui est requis étant donné le mode opératoire de notre processus de décision. Plusieurs étapes sont nécessaires à l'exécution de la méthode TOPSIS, détaillées dans les sous-parties suivantes.

Construction de la matrice - Soit m alternatives a et n attributs c pour chacune d'entre elles. Le regroupement de ces alternatives donne une matrice $X = \{x_{ij}\}$ pour $\{i \in \mathbb{N} \mid 1 \leq i \leq m\}$ et $\{j \in \mathbb{N} \mid 1 \leq j \leq n\}$.

Normalisation de la matrice et application des poids - Normaliser les critères ayant des unités et échelles différentes entre eux est nécessaire afin de pouvoir les comparer entre eux. C'est également à cette étape que nous appliquons les poids provenant des préférences de l'utilisateur.

$$v_{ij} = r_{ij} * \omega_j = \frac{x_{ij}}{\sqrt{\sum_{i=1}^{m} x_{ij}^2}} * \omega_j \tag{1}$$

Calcul des solutions idéales positive et négative puis mesure de l'écart avec chacune des alternatives - En sélectionnant les meilleures et les pires performances de chacun des critères de la matrice de décision normalisée pondérée, on peut déterminer les solutions idéales positive A^+ et négative A^- afin de mesurer l'écart de chacune des alternatives avec ces deux solutions que l'on notera S^+ et S^-.

$$PourA^+ = (v_1^+, ..., v_j^+), \quad (2) \qquad PourA^- = (v_1^-, ..., v_j^-), \quad (4)$$

$$Si^+ \triangleq \sqrt{\sum_{j=1}^{m} (v_{ij} - vj^+)^2} \quad (3) \qquad Si^- \triangleq \sqrt{\sum_{j=1}^{m} (v_{ij} - vj^-)^2} \quad (5)$$

Calcul de la distance relative avec la solution idéale - Cette dernière étape permet de donner un score à chaque alternative, qui représente sa distance avec la solution idéale. L'ordonnancement de ces scores permet de définir la meilleure alternative possible par rapport aux alternatives données ainsi que des préférences de l'utilisateur.

$$Ci = \frac{Si^-}{Si^+ + Si^-} \tag{6}$$

3. Application à un cas d'étude de gestion de chaîne logistique

Afin de tester le processus de décision automatisé proposé, nous avons sélectionné une étude qui propose d'introduire un système blockchain à une chaîne logistique afin de permettre le partage de données entre les différents acteurs (Longo *et al.*, 2019). Dans cette partie, nous détaillerons le scénario proposé par l'étude citée, puis les différents attributs requis pour la blockchain à implémenter qui découlent de ce sujet afin

d'exécuter notre processus de décision automatisé. Enfin, nos résultats sont validés en utilisant un outil permettant de tester ses performances implémenté dans ce but.

3.1. Scénario "Big-Box"

La chaîne logistique modélisée dans cette étude est constituée d'un réseau de détaillants de la chaîne de magasins Big-Box, ainsi que de trois grossistes qui alimentent leurs magasins. Les détaillants Big-Box étant regroupés dans une même organisation, l'étude considère qu'il y a un partage de données en temps réel, transparent et fiable entre les magasins. Cependant, les détaillants sont tout de même en compétition, car ils opèrent dans une même région géographique et proposent tous les mêmes gammes de produits. Les clients arrivent au magasin et sélectionnent des produits ainsi que leurs quantités respectives. Si le stock du magasin permet de satisfaire la demande, le produit concerné est réservé dans la quantité demandée; sinon, une réservation partielle est proposée; la demande non satisfaite est utilisée pour calculer les réapprovisionnements. L'inventaire est fait avant l'ouverture des magasins; si une commande est nécessaire alors le détaillant peut choisir l'un des grossistes pour s'approvisionner, en prenant en compte le délai d'approvisionnement, la demande actuelle et la quantité instantanément disponible pour les produits souhaités. Si la quantité d'un produit possédée par un grossiste devait ne pas être suffisante pour tous les détaillants, alors celui-ci est partagé équitablement.

Dans ce contexte, partager la demande globale des différents détaillants entre les grossistes pourrait permettre de prédire plus facilement le stock à constituer pour répondre aux demandes des détaillants. Cependant, les acteurs de ce système restent en concurrence et ne se font donc pas confiance mutuellement. Les auteurs de l'étude proposent donc la mise en place d'une blockchain permettant d'y enregistrer des données liées à la chaîne d'approvisionnement (notamment la demande du marché) sous la forme de valeur de hachage, ainsi que les différents tiers ayant accès à ces données (s'ils sont autorisés par la blockchain, ils peuvent faire directement une requête pour obtenir ces données auprès du tiers qui les a enregistrées). La sauvegarde de cette valeur permettant d'attester de la véracité des données transmises entre tiers, ils peuvent dorénavant se faire confiance entre eux.

3.2. Exigences du client Big-Box

Pour pouvoir préconiser la blockchain à l'aide de notre processus de décision, il convient d'identifier les attributs de qualité ainsi que les exigences et préférences quant à ces attributs (Section 2). Cette sous-section aborde donc chacun des attributs de qualité système proposés précédemment et explique le choix de la valeur de chacun d'entre eux.

Sécurité - Les données stockées étant hachées, elles ne sont pas considérées comme sensibles, pas plus que l'identité des tiers qui est masquée par leur adresse. Il est donc possible d'utiliser une blockchain publique (ce qui est d'ailleurs le choix initial de

l'étude), sans chiffrement des données. Les permissions étant par ailleurs gérées à l'échelle du contrat intelligent, il n'est pas nécessaire d'avoir une blockchain supportant la gestion de permissions. Par déduction, ces propriétés n'étant pas importantes dans ce contexte, elles sont toutes marquées comme étant *Indifférent* dans notre tableau d'entrées.

Efficacité - Le système blockchain n'a pas besoin de supporter un débit minimal de transactions par seconde (que l'on différencie du nombre de transactions par seconde supportable soumises en entrée) ainsi qu'une latence particulière. Néanmoins, une latence faible pouvant être profitable à l'expérience utilisateur, nous avons tout de même choisi de la fixer à *Faiblement désirable*. Pour ce qui est de l'efficience énergétique, c'est une propriété particulièrement intéressante dans une optique de réduction de coûts. L'utilisation de blockchains publiques à algorithme de consensus lourds (tel que PoW) est très coûteux en énergie. Nous avons donc choisi la préférence *Tout à fait désirable* pour cette propriété.

Fiabilité - Les acteurs ne se faisant pas confiance entre eux, il est indispensable d'avoir un pourcentage de tolérance aux fautes byzantines, ce qui indique que le système est capable de fonctionner correctement pour un certain nombre de nœuds pouvant avoir un comportement adverse. Nous avons choisi un pourcentage d'au moins 33,3%, ce qui permet de garantir la bonne continuité du réseau blockchain pour un nombre de nœuds fautifs $f + 1 < \frac{n}{3}$, n étant le nombre de nœuds totaux constituant le réseau.

Fonctionnalités - Pour répondre aux objectifs du sujet défini, la blockchain doit être capable de prendre la forme d'un élément de stockage pour contenir les données des détaillants ainsi que de supporter l'administration de celles-ci, de facto par le biais de contrats intelligents. Ces deux attributs sont donc définis respectivement comme *Avancé* ainsi que *Requis*. Les autres fonctionnalités n'étant pas nécessaires, elles sont marquées *Indifférent*.

Utilisabilité - Enfin, le dernier attribut choisi est la courbe d'apprentissage : dans un contexte où la blockchain doit permettre d'économiser des coûts associés à la chaîne d'approvisionnement ainsi que de supporter une application de faible complexité, utiliser une technologie dont il est facile d'en apprendre les mécaniques peut être un atout. Nous avons choisi de le marquer *Désirable*.

3.2.1. Compilation des valeurs choisies

La compilation des valeurs de ces qualités système aboutit au tableau 3, nous permettant d'exécuter notre processus de décision automatisé dans ce contexte.

3.3. Résultats

L'exécution du processus automatisé élimine l'alternative Bitcoin, car elle ne permet pas le support de contrats intelligents, ainsi que l'alternative Hyperledger Fabric, car elle ne tolère pas les fautes byzantines. Nous obtenons ainsi deux matrices,

Tableau 3. Exigences du système blockchain souhaité pour le cas d'étude.

Attributs	Exigences	Valeur exigée	Préférences
Ouvert publiquement	Aucune		Indifférent
Permissions	Aucune		Indifférent
Encryption native des données	Aucune		Indifférent
Débit (tx/s)	Aucune		Indifférent
Latence (s)	Aucune		Faiblement désirable
Efficient en énergie	Aucune		Tout à fait désirable
Tolérant aux fautes byzantines	Requis	$\geq$33,33 %	Désirable
Contrats intelligents	Requis	Oui	Indifférent
Cryptomonnaies	Aucune		Indifférent
Élément de stockage	Requis	Avancé	Indifférent
Élément de calcul	Aucune		Indifférent
Élément gestionnaire de biens	Aucune		Indifférent
Connecteur logiciel	Aucune		Indifférent
Courbe d'apprentissage	Aucune		Désirable

l'une contenant les poids et l'autre les alternatives possibles (resp. Ethereum-PoW, Ethereum-PoA, et Corda). Sachant qu'un poids à 0 pour un attribut donné rend celui-ci insignifiant dans le calcul du score de chaque alternative, nous pouvons simplifier ces matrices pour les valeurs suivantes :

$$W = \begin{pmatrix} 0.25 \\ 0.75 \\ 0.5 \\ 0.5 \end{pmatrix} \quad (7) \qquad A = \begin{pmatrix} 180 & 10 & 1 \\ 0 & 1 & 1 \\ 0.5 & 0.33 & 0.33 \\ 0.4 & 0.4 & 0.8 \end{pmatrix} \quad (8)$$

S'en suit l'exécution de notre algorithme d'aide à la décision, qui propose les résultats suivants (Figure 4). Notre algorithme de décision considère donc l'alternative Ethereum-PoA comme étant la meilleure. En effet, son score obtenu est le plus proche de 1 (solution idéale positive) des trois alternatives.

Tableau 4. Résultat de l'exécution du processus de décision.

Alternative	Score
Ethereum, PoA	0.83124114
Corda, PBFT	0.71016139
Ethereum, PoW	0.28983861
Hyperledger Fabric, Raft	Disqualifiée
Bitcoin, PoW	Disqualifiée

3.4. Validation de la solution proposée

Nous avons montré dans la sous-section précédente que la solution la plus adaptée est Ethereum-PoA. Pour confirmer la pertinence de la solution, nous allons expéri-

menter la robustesse du réseau Ethereum par le biais d'un outil permettant de tester ses performances développé dans ce sens.

Pour cela, nous avons implémenté un contrat intelligent pour Ethereum qui, lorsque déployé sur la blockchain, permet de faire les opérations définies dans le scénario de chaîne logistique (sauvegarde de données hachées, administration des tiers autorisés à utiliser l'application). Nous avons ensuite déployé une blockchain Ethereum-PoA sur Grid'5000, qui est un banc d'essai flexible, de grande taille et configurable à souhait pour le support d'expériences de large échelle. Les noeuds composant la blockchain possèdent chacun un processeur Intel Xeon Gold 5220 (18 cores), 96 GiB de mémoire vive, deux SSD de 480GB et 960GB respectivement, et une bande passante de 2x25 Gbps. Le nœud chargé de piloter l'expérience par l'envoi de transaction possède les mêmes caractéristiques techniques. Chacun des nœuds utilise le client d'Ethereum Geth, configuré avec l'algorithme PoA Clique [4], un intervalle de génération de bloc laissé à la valeur par défaut de 5 secondes, et une taille de bloc non limitée.

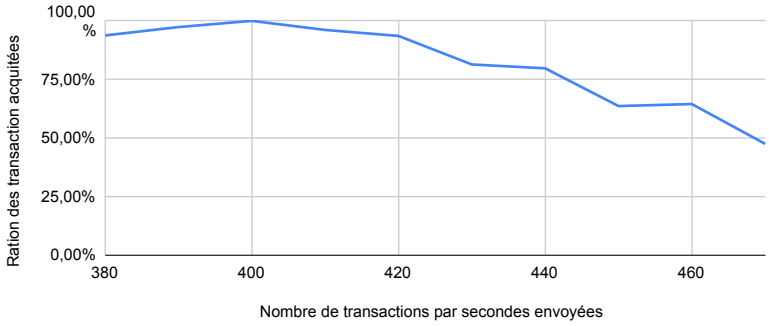


Figure 2. Tests de performance Ethereum-PoA pour le cas d'utilisation supply-chain Un client soumet à un réseau Ethereum-PoA contenant s noeuds, y transactions par seconde pendant n secondes. Après $n + \epsilon$ secondes (où ϵ étant égal à 3 fois la durée inter-blocs), le nombre z de transactions acquittées est retourné.

Nous constatons que la blockchain arrive bien à supporter une charge de 380 transactions par seconde. Nous en déduisons qu'une telle infrastructure est amplement capable de supporter une charge composée de 60 transactions par jour ainsi que quelques transactions ponctuelles d'administration du consortium. Le choix d'Ethereum-PoA est donc pertinent pour le cas d'utilisation donné.

4. Travaux connexes

Nos travaux s'inscrivent dans la lignée d'études réalisées pour faciliter l'adoption de la blockchain par le biais d'une aide à la décision entre différents types de block-

4. https://github.com/ethereum/EIPs/issues/225

chains, ou par la décision entre l'utilisation d'une blockchain ou non dans un contexte donné.

Dans (Wust, Gervais, 2018) les auteurs listent les principales propriétés de la blockchain (Transparence, intégrité, confiance ...) et proposent un modèle de décision sur l'adoption de la blockchain ou non en fonction de la réponse à certaines questions (telles que : "Y a-t-il plusieurs tiers impliqués ?" ou "Sont-ils de confiance ?") liées au cas d'étude donné. Ils appliquent ensuite leur modèle à plusieurs cas d'usages d'exemples. Bien qu'il y ait une étude des paramètres de la blockchain permettant de définir les questions du modèle de décision, le résultat est d'un niveau d'abstraction très élevé (blockchain publique, privée, permissionnée ou pas de blockchain). Il ne permet donc pas de prendre une décision précise sur la technologie de blockchain à utiliser ainsi que ces paramètres. Dans (Koens, Poll, 2018), les auteurs effectuent une revue de littérature sur des études relatives aux modèles de décision pour la blockchain afin de construire leur propre modèle. Les résultats de celui-ci sont un peu plus précis que le précédent, mais ne donnent toujours pas une recommandation précise. Les auteurs de (Labazova, 2019) ont également réalisé une revue de littérature tout en utilisant une approche DSR (Design Science Research) afin de construire leur modèle. Celui-ci comporte plusieurs niveaux de décision et prennent en compte des propriétés blockchain, ce qui permet à un utilisateur de faire un choix avec une précision accrue en sortie par rapport aux études précédentes. De plus, les auteurs montrent les dépendances entre certains paramètres (par exemple, la confidentialité et la transparence). Cependant, les paramètres en entrée sont majoritairement spécifiques à la blockchain et conditionnent l'utilisation du modèle par un expert. Une autre étude intéressante présente une troisième approche d'aide à la décision en proposant un travail complet de détail des fondamentaux blockchain dans la première partie de leur étude, ainsi qu'un modèle de décision introduisant des critères opposés (tels que performance/coûts), mais également une série de questions pour affiner le choix ("Quand utiliser la blockchain ?", "Quoi utiliser ?", "Comment utiliser cette blockchain ?") (Belotti *et al.*, 2019). Toutes ces études permettent de guider la prise de décision pour un projet blockchain donné, mais ne permettent pas d'aller plus en détail (paramètres blockchain) à cause des limitations des modèles de décision. Le manque d'automatisation et la résolution manuelle des questions ne permettent pas de prendre en compte un grand nombre d'exigences en entrée.

Certaines études ont été réalisées afin de répondre à cette problématique. À titre d'exemple, les auteurs de (Tang *et al.*, 2019) proposent d'utiliser une méthode d'aide à la décision multicritère appelée TOPSIS, qui est la même que celle utilisée dans cette étude, afin de déterminer la meilleure solution de blockchain publique disponible à partir d'un ensemble de critères en entrée. L'approche est intéressante dans ce contexte, mais ne permet pas de prendre en compte d'autres blockchains (privées, permissionnées). De plus, les critères techniques blockchain sont regroupés sous les critères "basic technology", "applicability" et "transaction per second", le premier étant quantifié via des experts, les recommandations données en résultat peuvent donc manquer de précision si l'on se place du point de vue de l'entreprise souhaitant démarrer son projet.

Dans (Farshidi *et al.*, 2020), les auteurs ont construit un système de prise de décision pour les technologies blockchain basées sur des travaux précédents pour d'autres technologies. Ils ont réalisé un sondage auprès d'experts pour déterminer les critères de choix les plus pertinents, puis ont rempli une base de connaissance contenant les valeurs de ces attributs choisis pour un ensemble large de blockchains (obtenus avec des livres blancs, études, tests de performance ...) afin de donner des recommandations via un moteur d'inférence. Leur outil est très performant et permet de donner des recommandations précises, nous voulons aller plus loin en proposant quelque chose plus orienté blockchain (prise en compte de processus métiers et de modèles architecturaux spécifiques) qui soit plus accessible pour des non-experts en blockchain, par le biais d'un modèle qui lie les attributs blockchain et qualité logicielle. Ainsi l'utilisateur peut saisir des exigences plus courantes que celles spécifiques à la technologie blockchain.

5. Discussion

La prédiction obtenue, qui est d'utiliser Ethereum-PoA, nous semble un choix pertinent pour plusieurs raisons. En effet, toutes les fonctionnalités que nous estimons nécessaires à la bonne implémentation du cas d'étude choisi sont présentes, tout en permettant de garantir un coût optimal de celle-ci (faible difficulté d'apprentissage et économique en énergie). Cependant, la méthode demeure sensible aux variations de poids. Si nous avions choisi un poids supérieur concernant le débit de transactions, nous aurions pu avoir un résultat différent en sortie. Des études de sensibilité peuvent permettre d'établir des intervalles, servant à indiquer à quel degré un poids peut varier sans affecter le résultat final. Il existe également des méthodes, comme celle de la détermination de poids par l'entropie, permettant de limiter l'impact des critères ayant une forte entropie en diminuant leur poids (Huang, 2008). Par ailleurs, l'échelle de Likert que nous avons choisi pour l'expression des préférences peut entraîner un biais selon la perception des écarts entre les différentes valeurs proposées par l'utilisateur. Afin de rendre le résultat plus fiable, d'autres systèmes de pondération pourraient être considérés (AHP).

Pour notre seconde expérience mettant en œuvre un test de performance de la blockchain Ethereum-PoA, nous avons trouvé que celle-ci n'était plus capable de traiter 100% des transactions entrantes à partir de 400 transactions par seconde. Le suivi de l'exécution sur chacun des nœuds montre que cette incapacité apparaît lorsque le CPU des noeuds n'est plus capable de supporter la charge de transactions reçues par le client Geth. Il est cependant possible de diminuer l'intervalle inter-blocs afin d'augmenter les performances, mais une valeur trop basse pourrait dégrader la qualité du réseau (difficulté à aboutir à un consensus entre nœuds d'autorité) et augmenter l'espace disque nécessaire (chaque bloc comportant au moins un entête de taille non nulle). Par conséquent, nous avons choisi de conserver la valeur par défaut, mais étudier l'impact d'une baisse sur la stabilité pourrait être profitable. Aussi, nous avons constaté lors de notre expérimentation que la courbe représente fidèlement la perte de transactions, mais nous pensons que répéter l'expérimentation pour chaque point

de mesure plusieurs fois et allonger le temps de chaque expérience pourrait affiner grandement les résultats.

6. Conclusion et travaux futurs

Dans cette étude, nous avons adapté une méthode d'aide à la décision multicritère afin de concevoir un processus de décision automatisé pour blockchain. Pour cela, nous avons sélectionné un panel pertinent de blockchains ainsi que de critères relatifs à la qualité d'un système (norme ISO 25010) pour créer une base de connaissance, puis nous avons choisi une liste de termes permettant à un utilisateur de soumettre ses préférences et exigences quant aux critères choisis pour la décision. Enfin, nous avons validé notre processus sur un cas d'étude de gestion de chaîne logistique et montré que notre outil est capable de recommander une blockchain alignée aux besoins de l'utilisateur. L'implémentation est en cours, et sera complétée puis mise à disposition en accès ouvert sur Github dans de futurs travaux. Cette étude est une première étape pour concevoir un processus de décision automatisé plus étendu, car il pourrait prendre en compte un plus grand nombre d'entrées (topologie d'architecture système, infrastructure, processus métiers...). Cela nous permettrait, à l'aide de ces informations, d'exécuter un test de performance personnalisé (tel que celui présenté dans la sous-section 3.4) pour chaque utilisateur avant même d'exécuter l'algorithme de décision, le but étant de fixer de manière extrêmement précise les valeurs des critères variants (débit de transactions, latence ...). Une autre piste d'amélioration est l'utilisation d'approches basées sur la logique floue ou les modèles bayésiens qui permettrait de tenir compte de l'aspect subjectif des critères de décision.

Remerciements

Les expériences présentées dans ce document ont été réalisées à l'aide du banc d'essai Grid'5000, soutenu par un groupe d'intérêt scientifique hébergé par INRIA et comprenant le CNRS, RENATER et plusieurs autres universités et organisations (voir https://www.grid5000.fr).

Bibliographie

Abeyratne S. A., Monfared R. P. (2016). Blockchain ready manufacturing supply chain using distributed ledger. *International Journal of Research in Engineering and Technology*, vol. 5, n° 9, p. 1–10.

Allen I. E., Seaman C. A. (2007). Likert scales and data analyses. *Quality progress*, vol. 40, n° 7, p. 64–65.

Androulaki E., Barger A., Bortnikov V., Cachin C., Christidis K., De Caro A. *et al.* (2018). Hyperledger fabric: a distributed operating system for permissioned blockchains. In *Proceedings of the thirteenth eurosys conference*, p. 1–15.

Belotti M., Bozic N., Pujolle G., Secci S. (2019). A Vademecum on Blockchain Technologies: When, Which and How. *IEEE Communications Surveys & Tutorials*, p. 1–1.

Brown R. G., Carlyle J., Grigg I., Hearn M. (2016). Corda: an introduction. *R3 CEV, August*, vol. 1, p. 15.

Ekblaw A., Azaria A., Halamka J. D., Lippman A. (2016). A case study for blockchain in healthcare:"medrec" prototype for electronic health records and medical research data. In *Proceedings of ieee open & big data conference*, vol. 13, p. 13.

Farshidi S., Jansen S., España S., Verkleij J. (2020). Decision support for blockchain platform selection: Three industry case studies. *IEEE Transactions on Engineering Management*, p. 1-20.

Herbaut N., Negru N. (2017). A model for collaborative blockchain-based video delivery relying on advanced network services chains. *IEEE Communications Magazine*, vol. 55, nº 9, p. 70–76.

Huang J. (2008). Combining entropy weight and topsis method for information system selection. In *2008 ieee conference on cybernetics and intelligent systems*, p. 1281–1284.

Hyvärinen H., Risius M., Friis G. (2017). A blockchain-based approach towards overcoming financial fraud in public sector services. *Business & Information Systems Engineering*, vol. 59, nº 6, p. 441–456.

Koens T., Poll E. (2018). What Blockchain Alternative Do You Need? In, p. 113–129. Springer.

Kornyshova E., Salinesi C. (2007). Mcdm techniques selection approaches: state of the art. In *2007 ieee symposium on computational intelligence in multi-criteria decision-making*, p. 22–29.

Labazova O. (2019). Towards a Framework for Evaluation of Blockchain Implementations. *ICIS 2019 Proceedings*.

Lai Y.-J., Liu T.-Y., Hwang C.-L. (1994). Topsis for modm. *European journal of operational research*, vol. 76, nº 3, p. 486–500.

Longo F., Nicoletti L., Padovano A., d'Atri G., Forte M. (2019). Blockchain-enabled supply chain: An experimental study. *Computers & Industrial Engineering*, vol. 136, p. 57–69.

Nakamoto S. (2008). *Bitcoin: A peer-to-peer electronic cash system.*

Szabo N. (1997, 9). Formalizing and securing relationships on public networks. *First Monday*, vol. 2, nº 9.

Takemiya M., Vanieiev B. (2018). Sora identity: secure, digital identity on the blockchain. In *2018 ieee 42nd annual computer software and applications conference (compsac)*, vol. 2, p. 582–587.

Tang H., Shi Y., Dong P. (2019). Public blockchain evaluation using entropy and TOPSIS. *Expert Systems with Applications*, vol. 117, p. 204–210.

Tapscott D. (2016). *Blockchain revolution: How the technology behind bitcoin is changing money, business, and the world*. Portfolio.

Wood G. *et al*. (2014). Ethereum: A secure decentralised generalised transaction ledger. *Ethereum project yellow paper*, vol. 151, nº 2014, p. 1–32.

Wust K., Gervais A. (2018). Do you need a blockchain? *Proceedings - 2018 Crypto Valley Conference on Blockchain Technology, CVCBT 2018*, nº i, p. 45–54.

Xu X., Pautasso C., Zhu L., Lu Q., Weber I. (2018). A pattern collection for blockchain-based applications. *ACM International Conference Proceeding Series*.

Recommandations basées sur les centres d'intérêts utilisateurs en Business Intelligence

Krista Drushku[1], Julien Aligon[2,3], Nicolas Labroche[2], Patrick Marcel[2], Verónika Peralta[2]

1. SAP Labs France
krista.drushku@sap.com

2. Université de Tours, LIFAT, (EA6300)
prenom.nom@univ-tours.fr

3. Université de Toulouse, IRIT, (CNRS/UMR 5505)
julien.aligon@irit.fr

RESUME. Cet article est un résumé des travaux proposés dans (Drushku et al. 2019) et publiés dans le journal « Information Systems », en 2019.

Mots-clés : Intérêt utilisateur, Construction d'attributs, Clustering, Analyse BI, Systèmes de recommandations collaboratifs

KEYWORDS: User interest, Feature construction, Clustering, BI analysis, Collaborative recommender systems

DOI:10.3166/RCMA.25.1-n © 2016 Lavoisier AR_DOI

De nos jours, il est assez courant pour des experts de domaine, analystes, cadres ou encore tout amateur de données, d'analyser de grands ensembles de données via des interfaces conviviales connectées à des systèmes de Business Intelligence (BI). Cependant, les systèmes BI actuels sont incapables de détecter et caractériser efficacement les intérêts des utilisateurs, ce qui peut conduire à des interactions fastidieuses et improductives. En particulier, une interaction avec un système BI peut être exprimée par l'utilisateur sous la forme d'une séquence de mots clés, traités ensuite par le système afin d'en déduire les requêtes formelles (généralement MDX ou SQL) les plus susceptibles d'être envoyées aux sources de données concernées (généralement des entrepôts de données ou des bases de données).

Ainsi, dans (Drushku et al. 2019), nous proposons un système de recommandation collaboratif spécifiquement conçu pour tirer parti d'intérêts utilisateurs. Ces intérêts utilisateurs sont déterminés par des clusters d'interactions

utilisateurs. Chaque interaction est représentée par un ensemble de descripteurs associés à des mesures de similarité. Le clustering repose ensuite sur une mesure de similarité qui est une composition linéaire des similarités locales par descripteurs. Cette similarité est apprise sous la forme d'un problème de classification binaire sur une base d'étiquettes fournies par des experts du domaine et indiquant si deux interactions sont liées au même intérêt utilisateur. Le détail de la détection des intérêts utilisateurs fait l'objet d'un précédent travail présenté dans (Drushku et al. 2017). En plus de ces intérêts identifiés automatiquement, nous proposons un système de recommandation collaboratif nommé *IbR* (**I**nterest-**b**ased **R**ecommender system). Le système *IbR* est basé sur un modèle de Markov construit à partir des clusters d'intérêts, représentant la probabilité pour un utilisateur de passer d'un intérêt à un autre.

Cette approche est validée expérimentalement par une étude utilisateur approfondie, à l'aide de traces réelles de navigation BI. Nos résultats sont doubles. Premièrement, nous montrons que notre mesure de similarité surpasse une mesure de similarité entre requêtes proposée dans l'état de l'art (Aligon et al. 2014) et donne une très bonne précision par rapport aux intérêts exprimés par les utilisateurs. Deuxièmement, nous comparons notre système de recommandation à deux systèmes de l'état de l'art, (Eirinaki et al. 2014) et (Aligon et al. 2015), démontrant l'avantage de s'appuyer sur les intérêts des utilisateurs.

Sur la base des résultats montrant la pertinence de notre approche, l'objectif à long terme est d'aller au-delà des systèmes d'interaction par mots clés. Par exemple, nous envisageons la mise en œuvre d'un assistant intelligent capable d'alerter l'utilisateur lorsque les sources de données sont actualisées ou lorsque les besoins d'informations des utilisateurs et leurs expertises changent. Dans ce but, nos travaux futurs comprennent le développement de systèmes de recommandation basés à la fois sur l'intérêt et les compétences des utilisateurs et leur validation via des études utilisateurs de plus grande ampleur.

Bibliographie

Drushku K., Aligon J., Labroche N., Marcel P., Peralta V. (2019). Interest-based recommendations for business intelligence users. Information Systems, Vol. 86, p. 79-93.

Drushku K., Aligon J., Labroche N., Marcel P., Peralta V., Dumant B. (2017). User interests clustering in business intelligence interactions. In Advanced Information Systems Engineering - 29th International Conference, CAiSE. p. 144-158.

Aligon J., Gallinucci E., Golfarelli M., Marcel P., Rizzi S. (2015). A collaborative filtering approach for recommending OLAP Sessions. Decis. Support Syst., Vol. 69, p. 20-30.

Aligon J., Golfarelli M., Marcel P., Rizzi S., Turricchia E. (2014). Similarity measures for OLAP sessions. KAIS, Vol. 39, p. 463-489.

Eirinaki M., Abraham S., Polyzotis N., Shaikh N. (2014). Querie: Collaborative database exploration. IEEE Trans. Knowl. Data Eng. 26, 1778– 1790.

Analyse de l'information dans les réseaux sociaux

Détection des attaques de confiance dans l'Internet des Objets Social - *Wafa Abdelghani, Florence Sèdes, Amel Corinne Zayani et Ikram Amous* (article long)

Détection d'événements géo-chrono-localisés sur Twitter - *Hosni Seffih, Myriam Lamolle, Aurélie Pradelles, Zhen Wang et Jérémy Lhez* (article long)

Analyse des discours sur Twitter dans une situation de crise - Étude de l'incident à l'usine Lubrizol de Rouen - *Hiba Jamra, Annabelle Gillet, Marinette Savonnet et Eric Leclercq* (article long)

Détection des attaques de confiance dans l'Internet des Objets Social

Wafa Abdelghani [1], **Florence Sèdes** [1], **Corinne Amel Zayani** [2], **Ikram Amous** [2]

1. IRIT, Université Paul-Sabatier, Toulouse, France
2. Miracl, Université de Sfax, Sfax, Tunisia

RÉSUMÉ. *L'Internet des Objets Social (SIoT) est un paradigme dans lequel l'Internet des Objets (IoT) est fusionné avec les réseaux sociaux. Dans ce type d'environnement, les participants sont en compétition afin d'offrir une variété de services attrayants. Néanmoins, certains d'entre eux ont recours à des comportements malveillants afin de propager des services de mauvaise qualité. Ils lancent ce qu'on appelle des attaques de confiance et brisent les fonctionnalités de base du système. Plusieurs travaux de la littérature ont traité ce problème et ont proposé différents modèles de confiance. Néanmoins, ces derniers proposent de classer les meilleurs noeuds du réseau SIoT. Ils ne permettent pas de détecter les noeuds malveillants. Pour remédier à ce problème, nous proposons un nouveau modèle de gestion de la confiance, capable de détecter et bloquer les noeuds malveillants afin d'obtenir un système fiable et résilient.*

ABSTRACT. *The Internet of Things Social (SIoT) is a paradigm where the Internet of Things (IoT) is merged with social networks. In this type of environment, participants compete to offer a variety of attractive services. Nevertheless, some of them resort to malicious behavior in order to spread poor quality services. They commit so-called trust-related attacks and break the basic functionality of the system. Several works in the literature have addressed this problem and have proposed different trust-models. Nevertheless, they propose to classify the best nodes of the SIoT network. They do not detect different types of trust attacks or malicious nodes. To address this problem, we propose a new trust evaluation model able to detect and block malicious nodes in order to ensure a reliable and resilient system.*

MOTS-CLÉS : *Internet des Objets Social, Réseaux sociaux, Gestion de la confiance, Attaques de confiance.*

KEYWORDS: *Social Internet of Things, Social Networks, Trust Management, Trust attacks.*

DOI:10.3166/HSP.x.1-16

1. Introduction

L'Internet des Objets (IoT) est dominé par un grand nombre d'interactions entre des milliards d'objets intelligents. L'intégration de la composante sociale dans l'Internet des Objets a donné naissance à l'Internet des Objets Social (SIoT). Le SIoT

est apparu suite à un processus évolutif qui a transformé les objets du quotidien en objets pseudo-sociaux capables d'interagir avec leur environnement, puis en objets sociaux, ayant la possibilité d'établir des relations avec d'autres objets, d'une manière autonome ((Atzori *et al*., 2014)). Cette nouvelle vision a permis de simplifier la navigabilité et la découverte des ressources ((Ali, 2015)), de garantir la scalabilité comme dans les réseaux sociaux classiques ((Atzori *et al*., 2012)) offrant une source de données plus riche et plus variée ((Geetha, 2016)). Dans ce type d'environnement, les participants sont en compétition afin d'offrir une variété de services attrayants. Néanmoins, certains d'entre eux ont recours à des comportements malveillants afin de propager des services de mauvaise qualité. Ils lancent ce qu'on appelle des attaques de confiance et compromettent les fonctionnalités de base du système.

Dans la littérature, la gestion de la confiance a été largement étudiée dans divers domaines. Plusieurs travaux se sont intéressés à ce problème et ont proposé différents modèles, basés sur différents facteurs et mesures. Notre contribution se résume comme suit: Contrairement à la majorité des systèmes de gestion de la confiance existants qui se limitent à classer les meilleurs noeuds du réseau, notre objectif est de détecter les noeuds malveillants. Cela permet de les isoler et d'obtenir un système de confiance. Pour ce, nous proposons un modèle de confiance basé sur de nouveaux facteurs qui sont dérivés de la description de chaque type d'attaque. Nous proposons, également, via l'apprentissage supervisé, de combiner les différents facteurs proposés afin de distinguer les comportements malveillants de ceux qui sont légitimes.

Le papier est organisé comme suit. Dans la section 2, nous présentons une étude des travaux de la littérature qui s'intéressent à la gestion de la confiance. Dans la section 3, nous détaillons le modèle de confiance proposé. Dans la section 4, nous présentons les expérimentations qui nous ont permis de prouver la résilience du modèle d'évaluation de la confiance. Enfin, nous concluons en section 5 et indiquons nos perspectives.

2. Concepts de base

L'internet des Objets Social permet aux personnes et aux objets d'interagir dans un cadre social pour soutenir un nouveau type de navigation. La structure du réseau SIoT peut être façonnée selon les besoins afin de faciliter la navigabilité, permettre la découverte d'objets et de services et garantir la scalabilité comme dans les réseaux sociaux humains. Toutefois, la confiance doit être assurée pour tirer parti des avantages multiples de ce paradigme.

La confiance est un concept complexe utilisé dans divers contextes et influencé par de nombreuses propriétés mesurables et non mesurables telles que la croyance, la fiabilité, l'intégrité ou encore l'aptitude. Il n'y a pas de définition consensuelle de ce concept. En effet, bien que son importance soit largement reconnue, les multiples approches pour la définition de la confiance ne se prêtent pas à l'établissement de mesures et de méthodologies d'évaluation.

La confiance peut être définie comme la croyance d'une entité en une autre pour accomplir un objectif selon ses attentes. Dans l'environnement SIoT, les entités peuvent être des êtres humains, des dispositifs, des systèmes, des applications ou encore des services. La mesure de la confiance peut être absolue (par exemple, la probabilité) ou relative (par exemple, un degré de confiance). L'objectif de la confiance peut être une action ou une information.

Différents modèles d'évaluation de la confiance sont proposés pour garantir la confiance dans différents types de systèmes. Leur rôle consiste à fournir (calculer) un score de confiance, qui aidera les acteurs à prendre la décision d'invoquer ou non les services fournis par d'autres participants. Il existe plusieurs attaques qui sont conçues pour briser spécifiquement cette fonctionnalité. Nous présentons dans cette section les principales attaques de confiance citées dans la littérature (Bao *et al*., 2013 ; R. Chen *et al*., 2016 ; Abdelghani *et al*., 2016).

2.1. Les attaques de confiance dans les réseaux SIOT

Une attaque est un comportement malveillant lancé sciemment par un noeud pour détruire, bloquer ou dégrader les fonctionnalités de base d'un système. Les attaques de confiance représentent un sous-ensemble des attaques possibles dans les environnements IoT et SIoT. Dans ce type d'attaques, un noeud malveillant peut promouvoir sa propre réputation pour accéder à des fonctions supérieures ou perturber le système de manière à réduire son efficacité. Ainsi, un dispositif IoT malveillant (sous contrôle d'un propriétaire malveillant) peut effectuer les attaques suivantes.

– **Bad Mouthing Attacks (BMA)** : est une attaque dans laquelle des noeuds malveillants tentent de détruire la réputation des noeuds bienveillants (en leur donnant de mauvais votes) afin de diminuer leurs chances d'être sélectionnés comme fournisseurs de services.

– **Ballot Stuffing Attacks (BSA)** : est une attaque dans laquelle des noeuds malveillants tentent de promouvoir la réputation d'autres noeuds malveillants afin d'augmenter leurs chances d'être sélectionnés comme fournisseurs de services.

– **Self Promoting Attacks (SPA)** : est une attaque dans laquelle des noeuds malveillants, fournissant des services de mauvaise qualité, tentent de renforcer leur réputation (en s'octroyant des votes élevés) afin d'être sélectionnés comme fournisseurs de services.

– **Discriminatory Attacks (DA)** : est une attaque dans laquelle des noeuds malveillants s'attaquent à d'autres noeuds qui ne présentent pas de relation sociale forte avec eux.

Dans le tableau 1, nous proposons une spécification informelle du comportement malveillant pour chaque type d'attaque de confiance.

Dans les différents types d'attaques, c'est le noeud qui invoque un service et l'évalue ensuite qui est malicieux. En effet, tous les types d'attaques de confiance opèrent par le biais de votes erronés et non représentatifs. Ce noeud malicieux (dit invoca-

Tableau 1. Spécification informelle du comportement malveillant pour chaque type d'attaque de confiance.

	Invocateur(u_i)	**Fournisseur**(u_j)	**Interaction**(u_i, u_j)
BMA	Noeud Malicieux: - Mauvaise réputation - Services de mauvaise qualité	Noeud bénin: - Bonne réputation - Service de bonne qualité	- Grand nombre d'interactions - Majorité de votes négatifs
BSA		Noeud Malicieux: - Mauvaise réputation - Services de mauvaise qualité	- Grand nombre d'interactions - Majorité de votes positifs
SPA		Noeud Malicieux: - Mauvaise réputation - Services de mauvaise qualité	- Grand nombre d'interactions - Majorité de votes positifs - Similarité
DA	Noeud Malicieux: - Mauvaise réputation	Noeud Malicieux/ Noeud bénin	Majorité de votes négatifs

teur) est caractérisé par une mauvaise réputation dans le réseau et par des services de mauvaise qualité. En effet, ce sont ces deux caractéristiques qui font qu'il ne parvient pas à propager ses services d'une manière légitime et a recours aux attaques de confiance pour le faire. Néanmoins, dans l'attaque BMA, l'invocateur va cibler un autre utilisateur (fournisseur de service) légitime, bien réputé et offrant des services qualifiés. Dans les attaques BSA et SPA, par contre, le noeud malicieux va cibler un autre noeud malicieux (lui-même dans le cas de SPA), dans l'objectif de s'entraider. Par contre, dans l'attaque DA, le noeud malicieux choisit ses cibles de manière aléatoire, sans se soucier du fait qu'elles soient légitimes ou malicieux. De ce fait, dans cette attaque, nous ne trouverons pas un grand nombre d'interactions avec un noeud donné. Or, dans les attaques BMA, BSA et SPA, le noeud malicieux va s'acharner sur une cible donnée, ce qui se reflète par un grand nombre d'interactions avec cette dernière. Enfin, dans les attaques BMA et DA, nous retrouvons une majorité de votes négatifs, car les noeuds invocateurs ont pour objectif de ruiner la réputation d'autre noeuds. Or, dans l'attaque BSA et SPA, les noeuds malicieux cherchent à promouvoir la réputation d'autres noeuds malicieux, engendrant une majorité de votes positifs.

2.2. Evaluation et gestion de la confiance

Les mécanismes de gestion de la confiance (MGC) permettent d'assurer le processus d'établissement, de propagation et de mise à jour de la confiance (Guo *et al.*, 2017). La figure 1 montre les différentes étapes d'un MGC.

L'étape d'établissement de la confiance se base sur "un modèle d'évaluation de la confiance" qui est construit en deux étapes. **(i) L'étape de composition** consiste à

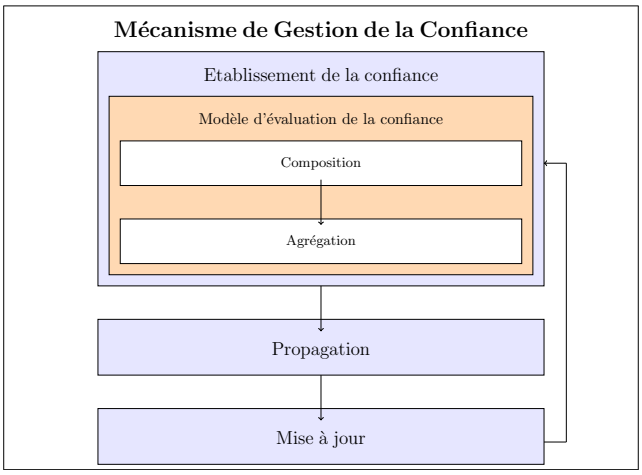


Figure 1. Architecture d'un mécanisme de gestion de la confiance

sélectionner les facteurs à prendre en compte dans le calcul des valeurs de confiance. Plusieurs facteurs ont été proposés dans la littérature, tels que l'honnêteté, la coopération, la similarité des profils, la réputation,... Ces derniers peuvent être classés selon différentes dimensions : (i) globaux ou locaux ; (ii) implicites ou explicites ; (iii) symétriques ou asymétriques. Pour les mesurer, les auteurs utilisent des informations relatives aux noeuds, telles que leur localisation ou leur historique d'interaction. **(ii) L'étape d'agrégation** consiste à choisir une méthode pour agréger les valeurs des différents facteurs afin d'obtenir la valeur de confiance finale. A cette fin, les auteurs de la littérature utilisent la moyenne pondérée, la logique floue, les modèles probabilistes, etc.

L'étape de propagation consiste à choisir une méthode pour propager dans le réseau les valeurs de confiance obtenues après l'étape d'agrégation. Deux méthodes sont utilisées. Dans la méthode dite centralisée une entité centrale fait les différents calculs pour tous les noeuds du réseau. Dans la méthode dite décentralisée, chaque noeud fait ses propre calculs. Certains travaux de la littérature utilisent la méthode de propagation centralisée, arguant que les noeuds impliqués dans les réseaux SIoT ont une capacité limitée (en termes de de calcul, de stockage, etc. ...). D'autres optent pour une approche décentralisée afin d'améliorer la scalabilité du système face à la montée en échelle (grand nombre de noeuds impliqués).

L'étape de mise à jour consiste à choisir une méthode pour mettre à jour les valeurs de confiance. Deux méthodes sont utilisées. Dans la méthode dirigée par le temps, les mises à jour sont faites d'une manière périodique. Dans la méthode dirigée par les évènements, les mises à jour se font à chaque fois qu'un nouvel évènement se produit.

Les étape de propagation et de mise à jour n'affectent pas la pertinence du MGC en termes de résilience face aux attaques. Cependant, elles ont un impact direct sur la performance du système. Nous nous concentrons dans ce travail sur l'étape principale

qui est celle de l'établissement de la confiance. En effet, la performance du système de gestion de la confiance dépend essentiellement du modèle mis en place pour évaluer le degré de confiance qui peut être accordé aux différentes entités impliquées dans le système.

3. Scénario de motivation

Prenons l'exemple d'un scénario dans le domaine des réseaux véhiculaires. Dans ce dernier, les conducteurs collaboreront pour connaître l'état de la route ou ils circulent. Les informations sur l'état de la route (accident, embouteillage, impasse, travaux, inondations, secousses, route étroite, etc.) peuvent être détectées automatiquement par différents types d'objets intelligents (véhicules intelligents, téléphones intelligents, capteurs, ...), ou signalées manuellement par différents conducteurs. Dans ce type de scénario, les services fournis sont les informations sur l'état d'un itinéraire donné à un moment donné. Une requête dans ce scénario se réfèrent à la position actuelle du conducteur en termes de longitude et de latitude et à un Δ_t se réfère à l'intervalle de temps actuel. Un conducteur qui emprunte un itinéraire donné lancera donc sa requête. Le système fonctionnera de manière à lui donner une réponse fiable. Dans un tel scénario, des attaques peuvent être effectuées pour différentes raisons. Certains conducteurs peuvent s'amuser à signaler des incidents (attaque discriminatoire). D'autres conducteurs peuvent s'entraider pour signaler un incident juste pour libérer le trafic sur leur trajet en effectuant des atttaques de type BSA. Mais d'autres raisons plus graves peuvent être à l'origine de ces les attaques telles que un vol ou un enlèvement à l'aide des attaques de type BMA.

4. Travaux connexes

Les modèles d'évaluation de la confiance se composent de deux étapes, à savoir (i) **l'étape de composition** et (ii) **l'étape d'agrégation de la confiance**. Le tableau 2 présente les facteurs proposés dans la littérature pour l'étape de composition. Ces facteurs représentent des concepts abstraits visant à quantifier le niveau de confiance des noeuds et sont calculés par différentes mesures en fonction de l'objectif et du contexte de l'auteur. Par exemple, dans (Jayasinghe *et al*., 2016), le facteur *recommandation* est mesuré comme le nombre de noeuds directement connectés à un noeud donné u_i. Toutefois, dans (Truong *et al*., 2016), le facteur *recommandation* est mesuré comme la moyenne totale des votes donnés à un noeud u_i. Cette même mesure (moyenne des votes) est appelée *réputation* dans certains autres ouvrages. Le facteur *coopération* est considéré comme un indicateur pour mesurer la connaissance d'un noeud dans (Truong *et al*., 2016) et est calculé comme la fréquence des interactions sociales entre deux noeuds. Toutefois, dans (R. Chen *et al*., 2016), le facteur *coopération* est calculé comme le nombre d'amis communs entre deux noeuds.
Etant donné qu'il n'existe pas de consensus sur la définition du concept de confiance, et compte tenu de la divergence des facteurs proposés, ainsi que des mesures proposées pour chaque facteur, nous avons choisi dans ce travail de partir de la définition

de chaque type d'attaque. En effet, nous estimons qu'un modèle d'évaluation de la confiance doit avant tout remplir le rôle de garant de la fiabilité du système dans lequel il est impliqué. Cette fiabilité est compromise par les différents types d'attaques de confiance.
Nous pensons que certains facteurs et mesures proposés dans la littérature, tels que le nombre d'amis communs ou le nombre de relations dans le réseau, n'ont aucun rapport avec les attaques de confiance citées. Il est, effectivement, courant (comme dans les réseaux sociaux classiques) qu'un noeud malveillant augmente le nombre de ses relations avant de procéder à des attaques. D'autres mesures, telles que la moyenne des votes reçus, pourraient donner une idée de l'historique d'un noeud et pourraient donc permettre de détecter certains types d'attaques. Les facteurs proposés dans la littérature restent insuffisants pour détecter tous les types d'attaques. En effet, aucun facteur ne permet, par exemple, de détecter l'attaque SPA dans laquelle un noeud est caché sous une fausse identité.

Pour conclure, la performance d'un modèle d'évaluation de la confiance dépend principalement des facteurs et des mesures choisies dans la phase de composition. Néanmoins, elle dépend également de la méthode choisie dans la phase d'agrégation. Le tableau 2 montre que la moyenne pondérée est la méthode d'agrégation la plus utilisée. Cependant, les comportements réalisés pour chaque type d'attaque de confiance ne sont pas similaires. Une moyenne pondérée ne peut pas détecter tous les types d'attaques car les facteurs considérés et les poids attribués à chaque facteur peuvent différer d'un type d'attaque à l'autre. En effet, prenons le cas de l'attaque SPA, le facteur similarité qui permet de détecter que c'est le même utilisateur sous une fausse identité est primordial. Alors qu'il n'a aucune importance dans le cas des attaques BMA, BSA et DA.
Le deuxième critère de comparaison concerne la résilience aux attaques de confiance. Certains des travaux cités s'intéressent à la détection des attaques (Z. Chen *et al.*, 2016 ; R. Chen *et al.*, 2016). Cependant, ils ne permettent pas de détecter tous les types d'attaques. Le tableau 2 montre que la majorité des travaux connexes proposent des modèles permettant d'attribuer un degré de confiance à chaque noeud du réseau (Truong *et al.*, 2017 ; Huang *et al.*, 2016 ; Militano *et al.*, 2016). Ces modèles proposent de classer les meilleurs noeuds du réseau en fonction de leurs valeurs de confiance. Leurs objectif est de recommander les meilleurs noeuds du réseau. Cependant, ce type de modèle ne permet pas de détecter et d'isoler les noeuds malveillants. Ceci leur donne libre accès pour établir différents types d'attaques dans le réseau. Le but de notre travail est d'isoler les noeuds malveillants afin d'obtenir un système fiable. Les noeuds jugés comme malveillants ne sont pas bloqués, mais seront naturellement moins sollicités, en raison.

5. Etape de composition: Sélection des facteurs

Dans cette section, nous présentons l'étape de composition de notre modèle d'évaluation de la confiance. Nous proposons de nouveaux facteurs permettant de décrire et de quantifier les différents comportements opérant dans les systèmes SIoT. Nos fac-

Tableau 2. Comparaison des travaux connexes

	Composition	Agrégation	Objectif
(Truong *et al*., 2017)	Connaissance Réputation Experience	LF	C
(Huang *et al*., 2016)	Consistence Intention Capacité	MP	C
(Truong *et al*., 2016)	Recommandation Réputation Experience	LF	C
(R. Chen *et al*., 2016)	Honnêteté Coopération Intêrets-communs	LC	DA
(Militano *et al*., 2016)	Fiablité Réputation	MP	C
(Z. Chen *et al*., 2016)	Réputation Relation Sociale Niveau d'énergie	MP	DA

teurs sont dérivés de la description informelle de chaque type d'attaque de confiance et permettent de distinguer les comportements malveillants des comportements bénins.

5.1. Réputation

Ce facteur représente la réputation globale d'un utilisateur u_i dans le réseau et est désigné par $Rep(u_i)$. Il est calculé comme le quotient entre le nombre d'interactions positives de u_i et le nombre total d'interactions (eq.1). Les interactions positives sont des interactions ayant reçu des valeurs de vote élevée. Les noeuds ayant une valeur de réputation élevée sont plus susceptibles d'être attaqués par d'autres noeuds. Les noeuds ayant une valeur de réputation faible sont plus susceptibles de lancer des attaques de confiance. Le facteur réputation, combiné à d'autres facteurs, permet de révéler des attaques de type BMA, BSA, SPA et DA.

$$Rep(u_i) = \frac{1}{N^i} \sum_{k=0}^{N^i} r_k \tag{1}$$

(Avec N^i est le nombre de votes attribués à l'utilisateur u_i et $r_k \in [0, 5]$ la valeur du vote .)

5.2. Honnêteté

Ce facteur permet d'estimer si un utilisateur est honnête et est désigné par Hon(u_i). Un utilisateur est considéré honnête si ses votes reflètent son opinion réelle, ce qui signifie qu'il n'essaie pas de donner des votes erronés dans l'objectif de promouvoir ou ruiner la réputation des autres utilisateurs. En effet, dans les attaques BMA, BSA et SPA, le noeud malveillant présente un comportement malhonnête. Dans l'attaque BMA, le noeud malveillant donne des valeurs de votes basses à un noeud qui fournit des services de bonne qualité, afin de ruiner sa réputation. Dans l'attaque BSA, le noeud malveillant donne des votes élevés à un autre noeud malveillant qui fournit des services de mauvaise qualité, dans le but de l'aider à promouvoir sa réputation. Dans l'attaque SPA, le noeud malveillant tente de promouvoir sa propre réputation en s'accordant de bons votes alors que ses services sont de mauvaise qualité.
La figure 2 montre comment ce facteur est calculé. Nous générons tous d'abord le vecteur de votes moyen $\bar{r}$ qui représente la moyenne des votes de tous les utilisateurs du système. Nous le comparons ensuite au vecteur de votes r_i de l'utilisateur u_i à l'aide de la similarité Cosinus.

$$
\begin{array}{c}
\begin{array}{cc} & \begin{array}{cccccccccc} u_1 & u_2 & \multicolumn{5}{c}{\ldots\ldots\ldots\ldots} & u_9 & u_{10} \end{array} \end{array} \\
\begin{array}{c} s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \end{array} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 \\ 4 & 5 & 5 & 4 & 4 & 5 & 5 & 4 & 5 & 4 \end{bmatrix} \\
\textit{Matrice de votes} \\
\textit{utilisateurs/ services}
\end{array}
\xrightarrow{Moyenne}
\begin{array}{c} \bar{r} \\ \begin{bmatrix} 0 \\ 0.5 \\ 1 \\ 5 \\ 4.5 \end{bmatrix} \end{array}
\xleftrightarrow[Cosinus]{Similarité}
\begin{array}{c} r_i \\ \begin{bmatrix} 5 \\ 5 \\ 5 \\ 0 \\ 1 \end{bmatrix} \end{array}
$$

Figure 2. Mesure du facteur "Honnêteté"

Le facteur Honnêteté peut simplement indiquer qu'un utilisateur a des opinions différentes, mais, associé à d'autres facteurs, il peut révéler différents types d'attaques. Pour mesurer et quantifier ce facteur, nous comparons le vecteur de votes de l'utilisateur $Rvec(u_i)$ avec la matrice de votes du réseau en utilisant la similarité cosinus (eq.2).

$$Hon(u_i) = MAX_r - \frac{1}{S}\sum_{j=0}^{S}\sqrt{(r_{i,j} - \bar{r}_j)^2} \tag{2}$$

(Avec $r_{i,j}$ la valeur du vote attribuée par l'utilisateur u_i au service s_j, $\bar{r}_j$ est la moyenne des votes attribués par tous les utilisateurs du réseau au service s_j, S est le nombre total de services et MAX_r est une variable statique indiquant la valeur maximale de vote.)

5.3. Qualité du fournisseur

La facteur Qualité du fournisseur permet de juger de la qualité des services fournis par un utilisateur donné. Il est désigné par $QoP(u_i)$. En effet, le noeud malveillant vise

à propager des services de mauvaise qualité. Les services de bonne qualité acquièrent naturellement une bonne réputation dans le réseau. Le noeud malveillant doit recourir à un comportement malveillant pour propager des services de mauvaise qualité et, par conséquent, il lance des attaques de type BMA, BSA, SPA et DA pour atteindre cet objectif. Le facteur QoP est donc essentiel pour distinguer les noeuds susceptibles d'opérer des comportements malveillants, des autres noeuds qui fournissent des services de bonne qualité et qui n'ont pas besoin d'avoir recours à des attaques pour les propager.

$$QoP(u_i) = \sum_{s_k \in S(u_i))} \alpha * QoS(s_k) + (1 - \alpha) * mr(s_k) \quad (3)$$

Avec S_{u_i} l'ensemble des services fournis par un utilisateur u_i, $QoS(s_k)$ est la valeur de QoS du service s_k, $mr(s_k)$ la moyenne des votes attribués à s_k et α est un poids.

5.4. Similarité

La similarité fait référence à la similitude entre l'utilisateur u_i et l'utilisateur u_j et est désignée par $SimU(u_i, u_j)$. Ce facteur est calculé sur la base de différentes caractéristiques telles que les profils, les intérêts, les services fournis, les dispositifs utilisés et la fréquence de proximité entre un couple d'utilisateurs. Elle vise à détecter les affinités entre les utilisateurs mais peut également révéler une attaque SPA dans laquelle le même utilisateur tente de promouvoir sa propre réputation sous une fausse identité.

5.5. Fréquence des votes

Ce facteur désigne la fréquence de votes attribués par un utilisateur u_i à un utilisateur u_j, et est noté par $RateF(u_i, u_j)$. Il est calculé comme le nombre de votes attribués par un utilisateur u_i à un utilisateur u_j divisé par le nombre total de votes donnés par l'utilisateur u_i. En effet, si un utilisateur u_i effectue une attaque contre un utilisateur u_j, nous trouverons probablement un grand nombre de votes attribués par l'utilisateur u_i à l'utilisateur u_j. Selon que ces votes soient positifs ou négatifs et en fonction de certains autres facteurs tels que la réputation et la qualité du fournisseur de l'utilisateur cible u_j, nous pouvons détecter une attaque de type BMA ou BSA.

5.6. Experience Directe

L'Expérience directe fait référence à l'opinion d'un noeud u_i sur ses interactions passées avec un noeud u_j, désignée par $ExpD(u_i, u_j)$. Elle est calculée comme le quotient des interactions réussies entre le noeud u_i et le noeud u_j, divisé par le nombre total d'interactions entre eux. Le facteur expérience directe ne peut donc pas révéler directement une attaque. Mais, combiné à d'autres facteurs, il permet de repérer le type de l'attaque. En effet, prenons l'exemple d'un noeud u_i qui attaque un noeud u_j. Ceci se traduit par des valeurs de réputation $Rep(u_i)$ et de Qualité de fournisseur $QoP(u_i)$

faibles pour u_i, ainsi que par une valeur de fréquence de votes $RateF(u_i, u_j)$ élevée qui reflète que u_i s'acharne à donner des votes au noeud u_j. Une valeur d'honnêteté de $Hon(u_i)$, vient confirmer l'hypothèse qu'il s'agit d'une attaque. Néanmoins, ces 4 facteurs combinés ensembles ne permettent pas de distinguer s'il s'agit d'une attaque BMA ou BSA. Le facteur Expérience directe permet de faire cette distinction. En effet, dans l'attaque BMA, le noeud u_i vise à ruiner la réputation de u_j et fournira donc des votes négatifs qui se traduiront par une valeur de $ExpD(u_i, u_j)$ faible, alors que, dans l'attaque BSA, le noeud u_i vise à promouvoir la réputation de u_j, ce qui donnera une valeur élevée de $ExpD(u_i, u_j)$.

5.7. Tendance des votes

Le facteur Tendance des votes est mesuré par le nombre de votes positifs divisé par le nombre total de votes fournis par un utilisateur. Elle vise à révéler si un utilisateur est plutôt optimiste ou pessimiste. Elle permet de détecter l'attaque discriminatoire (DA) dans laquelle l'utilisateur fournit des votes négatifs de manière aléatoire.

6. Etape d'agrégation: Conception de la fonction de classification

Une fois que nous avons choisi les différents facteurs qui permettent de décrire le comportement des utilisateurs du réseau, l'étape suivante consiste à choisir une méthode pour les agréger, afin d'obtenir la valeur de confiance finale. Dans la littérature, la méthode la plus courante est la moyenne pondérée. Cependant, nous estimons que la performance du système dépend, dans ce cas, principalement des poids attribués à chaque facteur. Ces derniers sont généralement fixés de manière empirique, or, l'importance des facteurs est subjective et dépend clairement des priorités de l'utilisateur. En outre, le comportement effectué pour chaque type d'attaque de confiance est différent. Une moyenne pondérée ne peut pas détecter tous les types d'attaques car les facteurs considérés et les poids attribués à chaque facteur différent d'un type d'attaque à un autre. En effet, si nous prenons l'exemple du facteur Similarité, ce dernier est essentiel pour détecter une attaque de type SPA, car il révélera qu'il s'agit du même utilisateur sous une fausse identité. Cependant, ce facteur n'a aucune importance dans le cas des attaques de type BMA, BSA ou DA.

La détection des noeuds malveillants étant considérée comme un problème complexe nécessitant une analyse approfondie du comportement des noeuds, nous proposons d'utiliser les techniques d'apprentissage automatique. Ainsi, nous considérons notre système comme un problème de classification. En effet, notre objectif est de détecter si un utilisateur est malveillant ou bénin. Un utilisateur est considéré comme malveillant s'il tente d'effectuer une attaque BMA, BSA, SPA ou DA. Si l'utilisateur n'a effectué aucune des attaques citées, il est considéré comme bénin. Ainsi, pour chaque couple d'utilisateurs (u_i, u_j), nous récupérons toutes les interactions passées. Nous calculons sur la base de ces interactions la valeur des différents facteurs liés à u_i, u_j et (u_i, u_j) (voir tableau 3). L'entrée de l'algorithme est l'ensemble de ces valeurs.

L'analyse de ces valeurs permettra de détecter si une attaque a eu lieu. En fonction de cela, l'utilisateur u_i sera jugé comme malveillant / bénin.

Tableau 3. Entrée de l'algorithme d'apprentissage automatique

$$\begin{matrix} u_i \\ \begin{bmatrix} Rep(u_i) \\ Hon(u_i) \\ QoP(u_i) \\ RateT(u_i,u_j) \end{bmatrix} \end{matrix} \quad \begin{matrix} (u_i,u_j) \\ \begin{bmatrix} SimU(u_i,u_j) \\ RateF(u_i,u_j) \\ ExpD(u_i,u_j) \end{bmatrix} \end{matrix} \quad \begin{matrix} u_j \\ \begin{bmatrix} Rep(u_j) \\ Hon(u_j) \\ QoP(u_j) \\ RateT(u_i,u_j) \end{bmatrix} \end{matrix}$$

7. Expérimentations et évaluations

7.1. Description du jeu de données et méthodologie

En raison du manque de données réelles, la majorité des travaux proposent des expérimentations basées sur des simulations. Dans notre travail, nous avons évalué les performances de notre modèle en nous basant sur des simulations appliquées à un jeu de données réel intitulé Sigcomm [1]. Ce dernier contient des utilisateurs, leurs profils, leurs listes d'intérêts, leurs relations sociales, leurs interactions et leurs localisations. Nous avons généré pour chaque utilisateur un ou plusieurs dispositifs et nous avons réparti ses interactions sur l'ensemble des dispositifs qui lui sont attribués. Nous avons considéré que 50% des utilisateurs de notre réseau sont malicieux et nous avons simulé pour chaque utilisateur malicieux l'une ou plusieurs des 4 types d'attaques de confiance décrits précedemment. Le tableau 4 présente les statistiques du data-set.

Tableau 4. Statistiques des données de test.

Contenu		Nombre
Utilisateurs	Malicieux	38
	Bénin	38
Profils	Institut Ville Pays	76
Intérêts		711
Relations sociales		531
Interactions		32000
Dispositifs		300
Services		364
Proximité		285788

Pour prouver la performance des facteurs proposés, nous avons mesuré le gain d'information pour chaque facteur séparément. Nous avons, ensuite, testé les diffé-

1. http://crawdad.org/thlab/sigcomm2009/20120715/

rents algorithmes d'apprentissage mis en oeuvre dans l'outil WEKA (Hall *et al*., 2009) pour construire notre modèle d'apprentissage. Ceci nous a permis de choisir la méthode d'apprentissage la plus adaptée à notre problématique. Enfin, afin de valider la méthode d'agrégation proposée (Apprentissage automatique supervisé), par rapport à la méthode d'agrégation la plus utilisée dans la littérature (Moyenne pondérée), nous avons comparé les résultats obtenus par (i) les autres travaux agrégés avec la moyenne pondérée, (ii) les facteurs que nous proposons, agrégés avec la moyenne pondérée et (iii) les facteurs que nous proposons, agrégés avec l'apprentissage automatique.

7.2. Sélection des facteurs et de l'algorithme d'apprentissage

Le gain d'information est une mesure d'évaluation utilisée pour sélectionner les attributs discriminiatifs et éliminer les attributs redondants, présentant des correlations ou inutiles. Elle se consiste à mesurer la variation de l'entropie en présence/absence d'un attribut (Azhagusundari, Thanamani, 2013 ; Lee, Lee, 2006 ; Yang, Pedersen, 1997). La figure 3 montre le gain d'information pour chaque facteur séparément. Le facteur *Similarité* (SimU) a la plus grande valeur de gain d'information. Cela s'explique par le fait qu'il est le seul facteur à permettre la détection les attaques de type SPA. Les facteurs *Fréquence de votes (RateF)*, *Qualité du fournisseur (QoP)*, *Tendance des votes (RateT)*, *Honnêteté (Hon)* et *Réputation (Rep)* présentent des valeurs de gain d'information presque égales. En effet, ils sont discriminatifs d'une manière égale pour la détection des attaques de type BMA, BSA et DA. Le facteur expérience directe a la plus faible valeur de gain d'information. En effet, ce facteur ne permet pas de détecter des attaques, mais, permet de faire la différence entre une attaque de type BMA et une attaque de type BSA. Nous avons ensuite

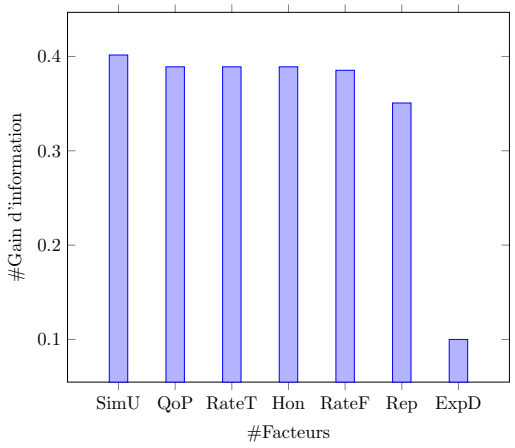


Figure 3. Gain d'information.

testé les différents algorithmes d'apprentissage mis en oeuvre dans l'outil WEKA ((Hall *et al*., 2009)) pour construire notre modèle d'apprentissage. Nous rapportons, dans Figure 4, les résultats obtenus pour les algorithmes : Naive Bayes, Multi-Layer Perceptron et Random Tree. Nous avons finalement opté pour le Multi-Layer Perceptron, car il a donné les meilleurs résultats en termes de F-Mesure.

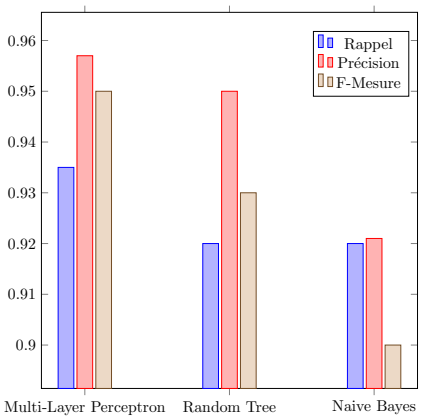


FIGURE *4. Comparaison des techniques d'apprentissage supervisé.*

7.3. Comparaison aux travaux connexes

Nous comparons les sept facteurs que nous proposons aux dix facteurs les plus utilisés dans la littérature. Nous avons implémenté ces dix facteurs pour les expérimenter sur notre jeu de données. Etant donné que les travaux de la littérature utilisent la moyenne pondérée pour agréger leurs facteurs, nous avons effectués différents essais pour fixer les poids et les seuils pour chaque travail. Nous avons ensuite utilisé la moyenne pondérée pour agréger les facteurs que nous proposons dans ce travail (F+MP). Ceci nous a permis de comparer et de valider la pertinence des facteurs proposés par rapport à ceux de l'état de l'art. Autrement dit, de valider notre proposition pour l'étape de composition indépendamment de la méthode d'agrégation que nous proposons ensuite. Enfin, nous avons appliqué notre méthode d'agrégation, notamment l'apprentissage automatique supervisé (F+AA), pour prouver sa pertinence par rapport à la méthode d'agrégation la plus utilisée dans la littérature (la moyenne pondérée). La figure 5 montre les résultats obtenus. Les facteurs proposés donnent de meilleurs résultats en termes de rappel, de précision et de F-mesure par rapport aux autres travaux, même dans le cas de l'agrégation avec la moyenne pondérée. Les résultats sont encore meilleurs lorsque nous appliquons la technique d'apprentissage automatique.

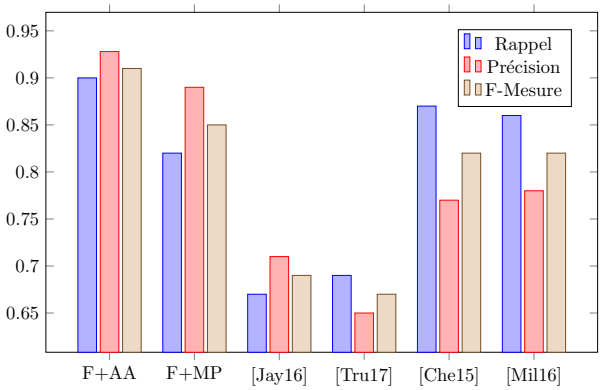


Figure 5. Comparaison avec les travaux connexes

8. Conclusion et perspectives

Nous avons proposé un modèle d'évaluation de confiance, capable de détecter les noeuds malveillants dans les environnements SIoT. Ce modèle repose sur de nouveaux facteurs, permettant de quantifier le comportement des utilisateurs, ainsi qu'une nouvelle méthode d'agrégation permettant d'analyser ces comportements. Des expérimentations basées sur des données réelles nous ont permis de prouver la pertinence du modèle proposé. Dans la suite de ce travail, nous nous intérressons aux étapes de propagation, de stockage et de mise à jour des valeurs de confiance qui sont primordiales pour la mise en oeuvre d'un mécanisme de gestion de la confiance adapté aux contraintes des environnements SIoT (scalabilité, dynamisme, minimisation de la consommation des ressources...).

Bibliographie

Abdelghani W., Zayani C. A., Amous I., Sèdes F. (2016). Trust management in social internet of things: A survey. In *Social media: The good, the bad, and the ugly*, p. 430–441. Swansea, Springer.

Ali D. H. (2015). *A social internet of things application architecture: applying semantic web technologies for achieving interoperability and automation between the cyber, physical and social worlds*. Thèse de doctorat non publiée, Institut National des Télécommunications.

Atzori L., Iera A., Morabito G. (2014). From" smart objects" to" social objects": The next evolutionary step of the internet of things. *IEEE Communications Magazine*, vol. 52, n° 1, p. 97–105.

Atzori L., Iera A., Morabito G., Nitti M. (2012). The social internet of things (siot)–when social networks meet the internet of things: Concept, architecture and network characterization. *Computer networks*, vol. 56, n° 16, p. 3594–3608.

Azhagusundari B., Thanamani A. S. (2013). Feature selection based on information gain. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, vol. 2, nº 2, p. 18–21.

Bao F., Chen I., Guo J. (2013). Scalable, adaptive and survivable trust management for community of interest based internet of things systems. In *11th international symposium on autonomous decentralized systems*, p. 1–7. Mexico City, IEEE.

Chen R., Bao F., Guo J. (2016). Trust-based service management for social internet of things systems. *IEEE transactions on dependable and secure computing*, vol. 13, nº 6, p. 684–696.

Chen Z., Ling R., Huang C., Zhu X. (2016). A scheme of access service recommendation for the social internet of things. *Int. J. Communication Systems*, vol. 29, nº 4, p. 694–706.

Geetha S. (2016). Social internet of things. *World Scientific News*, vol. 41, p. 76.

Guo J., Chen R., Tsai J. J. (2017). A survey of trust computation models for service management in internet of things systems. *Computer Communications*, vol. 97, p. 1–14.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H. (2009). The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, vol. 11, nº 1, p. 10–18.

Huang J., Seck M. D., Gheorghe A. (2016). Towards trustworthy smart cyber-physical-social systems in the era of internet of things. In *System of systems engineering conference (sose), 2016 11th*, p. 1–6. Kongsberg, Norway, IEEE.

Jayasinghe U., Truong N. B., Lee G. M., Um T.-W. (2016). Rpr: A trust computation model for social internet of things. In *Ubiquitous intelligence & computing, advanced and trusted computing, scalable computing and communications, cloud and big data computing, internet of people, and smart world congress*, p. 930–937. Toulouse, France, IEEE.

Lee C., Lee G. G. (2006). Information gain and divergence-based feature selection for machine learning-based text categorization. *Information processing & management*, vol. 42, nº 1, p. 155–165.

Militano L., Orsino A., Araniti G., Nitti M., Atzori L., Iera A. (2016). Trusted d2d-based data uploading in in-band narrowband-iot with social awareness. In *Personal, indoor, and mobile radio communications (pimrc), 2016 ieee 27th annual international symposium on*, p. 1–6. Valencia, Spain, IEEE.

Truong N. B., Um T.-W., Lee G. M. (2016). A reputation and knowledge based trust service platform for trustworthy social internet of things. *Innovations in Clouds, Internet and Networks (ICIN), Paris, France*.

Truong N. B., Um T.-W., Zhou B., Lee G. M. (2017). From personal experience to global reputation for trust evaluation in the social internet of things. In *Globecom 2017-2017 ieee global communications conference*, p. 1–7. Singapore, IEEE.

Yang Y., Pedersen J. O. (1997). A comparative study on feature selection in text categorization. In *Icml*, vol. 97, p. 35.

Détection d'événements géo-chrono-localisés sur Twitter

Hosni Seffih [1, 2], **Myriam Lamolle** [2], **Aurélie Pradelles** [1], **Zhen Wang** [1], **Jérémy Lhez** [1]

1. *GEOLSemantics, 12 avenue Raspail, 94250 Gentilly*
   prenom.nom@geolsemantics.com
2. *LIASD - IUT de Montreuil - Université Paris 8, 140 Rue de la Nouvelle France, 93100 Montreuil*
   m.lamolle@iut.univ-paris8.fr

RÉSUMÉ.
Dans cet article, nous présentons un processus complet d'extraction d'événements géo-chrono-localisés dans des tweets sous forme de triplets (temps, lieu, action), mais aussi une technique de regroupement de tweets parlant d'un même événement afin d'apporter au fur et à mesure du processus de détection plus de précision sur la date et le lieu de l'événement. À cette fin, notre système applique une analyse linguistique généraliste fondée sur des règles linguistiques ainsi qu'une extraction sémantique guidée par une ontologie d'événements. Ces deux phases ont été adaptées pour l'extraction d'information dans le contexte de Twitter où les messages sont courts, les abréviations et fautes d'orthographe nombreuses. Un échantillon de tweets collectés récemment sera testé sur notre système et sur un système d'apprentissage statistique (i.e. Machine Learning).

ABSTRACT.
In this paper, we present the complete process of extracting geo-chrono-localized events from tweets. We have designed a system capable of detecting an event by a triplet (time, place, action), and the first published tweet concerning it if possible. We also show how tweets are grouped talking about the same event in order to bring more precision during the detection process of event date and location. Our system applies general linguistic analysis based on linguistic rules, and a semantic extraction guided by an events ontology. The two phases have been adapted for information extraction in the context of Twitter where messages are short and contain many abbreviations and misspellings. In order to demonstrate the effectiveness of our approach, a sample of tweets will be tested on our system and a machine learning system.

MOTS-CLÉS : *Twitter, Système d'information, Détection d'événements géo-chrono-localisés, Regroupement de tweets, Ontologie, Machine learning*

KEYWORDS: *Twitter, Information system, Detection of geo-chrono-localized events, tweets' merging, Ontology, Machine learning*

1. Introduction

De façon simplifiée, un système d'information (SI) peut être vu comme un ensemble de sources de données et d'applications permettant la gestion d'information. Avec l'avènement du Web et de l'*open data*. Twitter représente une des sources externes, très utilisées, mais nécessitant une adaptation des ressources propres au SI traitant habituellement des textes de grande taille pour rechercher des informations particulières ; d'autant plus lorsque une fonctionnalité du SI est la détection d'événements à des fins d'alerte de danger (par exemple, un feu, un tremblement de terre, etc.).

La détection d'événements « géo-chrono-localisés » n'est pas chose facile sur Twitter sachant qu'elle doit se faire au plus tôt donc en un minimum de tweets (Ying *et al*., 2018). De façon générale, beaucoup d'ambiguïtés peuvent survenir des termes utilisés s'ils ne sont pas replacés dans leur contexte. Par exemple, le terme « Nice » peut faire référence à la ville du sud de la France mais aussi à la notion de « joli » (selon l'anglais) alors que la conversation est en français. Autant nous sommes capables intuitivement de faire cette distinction et donc de localiser ou non l'événement, autant cela devient difficile pour une machine ou un programme. Dans le cas des réseaux sociaux, et plus précisément dans les tweets, le vocabulaire et la tournure des phrases sont encore plus succincts puisque contraints à 144 caractères jusqu'à récemment, en 2017 passage à 280 caractères (Perez, 2017). Par exemple, il apparaît beaucoup plus de mots contractés ("tkt" pour "Ne t'inquiète pas"), de phonétisations ("foto" pour "photo" ou "C" pour "c'est"), de pictogrammes, etc. Quant à la détection d'événements, nous pouvons être guidés par les *hashtags* mais cela ne nous permettra pas de déterminer où et quand se déroule(-era) l'événement. Ces informations sont essentielles lors du déclenchement d'alertes automatiques pour des risques climatiques, des mises en sécurité de personnes, etc., en temps réel ou semi-réel.

Cet article s'inscrit dans cette perspective pour, d'une part, extraire et améliorer le plus possible la sémantique d'un tweet, et, d'autre part, augmenter la pertinence des alertes par une datation et une précision de localisation acceptables. Concernant notre premier objectif, les outils d'extraction sont fondés sur une analyse morphosyntaxique profonde généraliste puis sur une extraction sémantique détectant les concepts décrits dans l'ontologie de l'application grâce à des règles d'extraction utilisant les résultats de l'analyse morphosyntaxique profonde. Afin d'affiner la pertinence du déclenchement de l'alerte, notre deuxième objectif, nous allons nous attacher à compléter une ontologie existante dont la modélisation va être dirigée par un corpus *ad hoc* au contexte des événements afin d'inférer des motifs (dits *patterns*) de « cause à effet », problème assez complexe selon (Besnard *et al*., 2008) qui ont présenté un ensemble de modèles d'inférence formels, depuis des déclarations causales jusqu'aux déclarations explicatives. Ces modèles présentent des prémisses ontologiques qui sont considérées comme essentiels pour déduire les déclarations explicatives. Le domaine d'application étant l'extraction d'événements demandant une intervention des autorités locales, nous limitons notre étude à des événements présents ou du passé très récent.

Dans le cadre du projet "Safecity", GEOLSemantics réalise un système capable de détecter dans les réseaux sociaux (en particulier Twitter) des événements (incendie, accidents, inondation, etc.) qui pourraient nécessiter l'intervention des autorités locales (pompier, ambulance, police, etc.), et de caractériser ces événements en particulier avec leur localisation précise. À la demande des partenaire, nous devons envoyer l'alerte via Kafka [1] dans un contexte de *Stream processing*.

L'article est structuré comme suit : après un état de l'art sur le traitement des tweets notamment pour la détection des événements et des lieux, nous allons présenter notre processus de détection d'événements. La section suivante décrit les différentes briques logicielles sur lesquelles repose notre processus. Il s'ensuit une section concernant l'expérimentation proprement dite dans laquelle nous comparerons la partie détection d'événement de notre système avec un système d'apprentissage statistique. Nous finirons par une conclusion générale et les perspectives envisagées.

2. État des lieux sur la détection d'événements

Twitte est un site de microblogging, qui a permis, entre autres, de démocratiser la diffusion de l'information (Khamis *et al*., 2017), même si cela comporte un risque de propagation de fausses nouvelles (Guibon *et al*., 2019). L'orthographe liée à l'écriture des messages s'est bien améliorée depuis les premiers tweets, mais il reste que ces messages sont souvent écrits phonétiquement, avec des mots collés et/ou des abréviations (140 signes passés à 280 depuis novembre 2017). Les outils d'analyse linguistique et d'extraction sémantique doivent être adaptés pour être robustes à ce type de rédaction. Cela se traduit par des interprétations de mots plus ambiguës et par une syntaxe plus simple avec un manque de mots-outils mais, cependant, dans un ordre plus standard et avec une construction simplifiée des phrases.

Les deux problèmes majeurs rencontrés lors de l'extraction d'événements à partir des tweets sont l'imprécision temporelle (la date/heure de l'événement n'est pas mentionnée systématiquement, ou des expressions de date/heure relatives sont utilisées comme « vient de », « va organiser », « demain », etc.) et l'imprécision géographique (« pas loin de », « devant », etc.).

Concernant la détection des dates dans les tweets, par exemple, pour la prédiction d'événements, en se basant sur une approche d'analyse sémantique automatique et sur des traitements NLP [2] sur des tweets, et en utilisant une modélisation linéaire, (Wang *et al*., 2012) ont pu prédire des délits de fuite. (Aramaki *et al*., 2011) ont pu démontrer que des instruments NLP peuvent permettre l'extraction d'informations pertinentes, en utilisant un classifieur SVM [3] sur un corpus de tweets sur la grippe. (Sakaki *et al*., 2010) ont pu détecter en temps réel des tremblements de terre et lancer une alerte en

1. https://kafka.apache.org/
2. Natural Language Processing
3. Support Vector Machine

avertissant des utilisateurs enregistrés, alerte émise bien avant celle de la JMA (Japan Meteorological Agency).

La détection des lieux, quant à elle, se fait sur quelques paramètres à savoir le tweet de l'utilisateur, son réseau (*followers*, *following*) et ses métadonnées (*profile location*, *tweet timezone*). Selon (Hecht *et al*., 2011), 34% des lieux déclarés sont des lieux sarcastiques. Twitter permet aux utilisateurs d'afficher leurs coordonnées GPS mais seulement 1% des utilisateurs le font vraiment selon (Han *et al*., 2013). Nous avons vérifié cet état de fait sur environ 1000 tweets extraits sur la région parisienne par *bounding box*, c'est-à-dire que les utilisateurs permettent à l'application de se connecter à leur GPS. Ce test a révélé que 19% des tweets n'avaient pas déclaré de lieu dans *profile location*, 38% ont déclaré un autre lieu, 16% ont déclaré un lieu inexistant (cas des *troll*) et seulement 28% ont déclaré la bonne localisation, des enseignes de commerce pour la plupart. D'autre part, nous avons aussi réalisé un autre test en extrayant 1000 tweets commençant par *"Je suis à Paris"*. Dans ce cas, 38% des utilisateurs n'avaient pas déclaré le lieu, 35% avaient déclaré un autre lieu, 20% étaient des trolls et seulement 8% avaient déclaré la bonne géolocalisation Paris. Le lieu déclaré dans le tweet est considéré comme une valeur sûre pour la détection du lieu par rapport à la géolocalisation de Twitter. Dernièrement, (Huang, Carley, 2019) a proposé une géolocalisation hiérarchique pour des utilisateurs de Twitter, fondé sur un réseau de neurones.

Ces deux détections sont essentielles pour détecter des événements, notamment lors de la gestion des catastrophes naturelles (Kryvasheyeu *et al*., 2015) ou le suivi d'une épidémie (Broniatowski *et al*., 2013) entre autres. Pour (Hoang, Mothe, 2018) les événements sont représentés par trois dimensions principales: le temps, le lieu et les informations relatives à l'entité, ils ont fait des travaux sur la relation rappel/précision, combinant différentes méthodes pour extraire les lieux. En ce qui nous concerne, nous visons essentiellement à identifier les événements contenant au minimum une indication de lieu et une temporalité au présent ou au passé proche, exprimés dans le contenu du message, afin de lancer des alertes. Les informations relatives à l'événement permettent de le caractériser par des précisions sur qui (personne ou organisation) fait quoi (action, événement, expérience), quand (date), où (lieu), avec quoi (objet) et combien (quantité ou mesure). Ceci devrait permettre de lever l'ambiguïté des entités et des actions grâce à la définition des contextes sémantiques.

D'autre part, un tweet ne peut pas toujours refléter à lui seul l'importance d'un événement. Il faut être capable d'identifier qu'un autre message relate le même événement ou non (Jackoway *et al*., 2011). C'est indispensable pour ne pas « noyer » l'alerte par des informations redondantes. Il faut aussi être capable de compléter l'information au fur et à mesure que de nouveaux messages arrivent. Enfin, en ce qui concerne des événements annoncés, il faut pouvoir les suivre pour confirmer s'ils arrivent ou non, s'ils auront lieu ou non. En suivant les méthodes orientées caractéristiques (*feature-pivot methods*) qui sont plus sémantiques car elles étudient la distribution des mots et extraient des événements selon un motif de mots défini (dit *pattern*) (Kleinberg, 2003), un événement précis pourrait être défini. Un événement est donc représenté par un

certain nombre de mots-clés avec un nombre d'apparition (Weng, Lee, 2011). Nous pouvons aussi observer que la détection d'un événement dans un lieu et une date précis nous incite à avoir une micro-détection de la date, du lieu et de l'action avant de passer à la macro-détection de l'événement.

3. Processus de détection d'événements

Nous allons maintenant aborder le processus qui permet de détecter chaque élément constitutif d'un événement E pour lancer une alerte, à savoir A = action, L = Lieu, D = Date. Ces trois éléments vont nous permettre de regrouper des tweets parlant du même événement, et ce, même si dans certains tweets, tous les éléments ne sont pas renseignés. Il s'agit donc d'obtenir : $E = (A, L, D)$ où $A = f_{action}(\{tweet\}_n)$, $L = f_{lieu}(\{tweet\}_n)$, $D = f_{date}(\{tweet\}_n)$ et f représente la fonction de raffinement d'un des critères $action$, $lieu$, $date$. Notons que $lieu$ peut être un lieu précis ou une zone géographique, et $date$ une date précise ou un intervalle de temps. La figure 1 présente l'architecture de notre approche constituée de trois parties : la collecte des tweets et l'extraction d'information en deux dimensions (partie de gauche), la chaîne de traitement GEOL en quatre étapes (partie centrale) et la sortie de l'événement formalisé pour l'envoi d'alerte en JSON (partie de droite) dont le détail est donné dans les sous-sections suivantes.

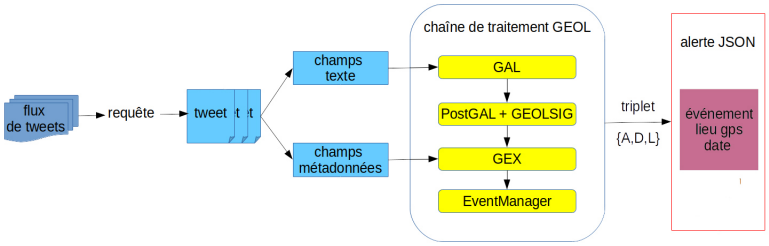


Figure 1. Processus complet de détection d'événements géo-chrono-localisés

3.1. Collecte des tweets

Il existe plusieurs solutions pour collecter des tweets. Citons :

– **l'utilisation de l'API Twitter :** nous pouvons interroger l'API officielle de Twitter pour collecter les tweets et un nombre considérable de métadonnées indisponibles sur l'interface Twitter. Avec la version gratuite, nous ne pouvons récupérer que 1% des tweets à un instant t (par l'API Stream) et filtrer les tweets selon des mots-clés, les coordonnées de géolocalisation et l'identifiant (ID) de l'utilisateur ou la langue de rédaction du tweet. Avec l'API REST, des tweets spécifiques peuvent être extraits, les tweets d'un utilisateur en particulier (les 3200 plus récents), ou par *bounding box* sur un espace géographique donné. Dans ce cas, seuls les tweets des utilisateurs qui autorisent l'application à se connecter à leurs coordonnées GPS sont récupérables. Notons

que dans la version gratuite le nombre de requêtes va de 180 à 450 demandes selon le type d'authentification toutes les 15 minutes [4] (Steinert-Threlkeld, 2018).

– **le *scrapping* :** cette approche consiste à concevoir un programme qui va interroger l'interface web de Twitter afin d'obtenir ensuite la page web de la réponse à la requête et d'en extraire les tweets en relation avec la requête et les métadonnées disponibles, en utilisant l'outil gratuit Twint [5].

Malgré la stabilité et la robustesse de l'API Twitter, compte-tenu des restrictions très fortes quant à son utilisation en version gratuite, nous avons décidé d'utiliser le *scrapping* par Twint car la quantité de tweets est bien supérieure et le texte collecté nous permet de constituer un corpus bien plus complet et pertinent qui sera utilisé par la chaîne de traitement GEOL (*cf.* figure 1).

3.2. Prétraitement des tweets

En plus des dictionnaires argotiques, GEOLSemantics a développé un outil permettant la détection de mots dans un grand flux de données grâce à un dictionnaire arborescent dont chacun des nœuds contient une table de hachage indexant ses branches. Le dictionnaire est d'abord fléchi pour obtenir toutes les formes de chacun des mots. Ce dictionnaire est ensuite phonétisé, puis regénéré en parsant le fichier d'entrée caractère par caractère (*cf.* 2). Le fait de changer les mots de l'argot vers le français et de corriger les fins de mots permet de constituer une entrée qui garantit une analyse sémantique plus propre.

```
Dictionnaire ayant déjà une entrée pour "poire" -> p-o-i-r-e-null
1. On rencontre dans le fichier de génération du dictionnaire le mot "poirier"
2. On suit les branches en cherchant chaque caractère: ROOT-p-o-i-r-
3. On arrive à "i" et on le cherche dans la table de #HashTable_de_poir
4. La table renvoie "null" le i n'existe donc pas, on génère une entrée dans la table pour le i
5. On génère successivement un nœud + une entrée dans ce nouveau nœud pour les caractères suivants
6. On change la valeur de "EndOfWord" a true pour le dernier "r" dans l'arbre poirier
7. On obtient donc l'arbre: ROOT-p-o-i-r-e-null
                                \
                                 i-e-r-null
```

Figure 2. Processus d'écriture de Poire *et ses déclinaisons dans le dictionnaire*

3.3. Extraction d'événements

Une fois le pré-traitement réalisé, lors de la phase d'extraction, nous avons cherché à repérer des événements composés d'une action, d'un lieu et d'une date. Nous avons vu précédemment que les tweets sont des textes courts, mal rédigés et avares en informations. Aussi, un certain nombre d'adaptations des traitements linguistiques couramment employés sur des documents textuels sont nécessaires. Cette extraction s'appuie principalement sur deux modules (*cf.* GAL et GEX de la figure 1) qui inter-

4. https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets
5. Twitter Intelligence Tool, https://github.com/twintproject/twint/wiki

agissent avec une ontologie propre à GEOLSemantics pour améliorer la sémantique de l'information à traiter.

3.3.1. Ontologie

Cette ontologie se fonde sur des logiques de description qui définissent le sens des concepts par leurs caractéristiques et par une représentation structurée ou formelle de leur rôle dans le domaine. Pour les besoins de nos travaux, nous nous sommes intéressés aux ontologies événementielles. Citons par exemple LODE (Troncy *et al.*, 2010), une ontologie pour représenter des événements dans le web des données, SEM qui permet d'identifier les événements passés, présents et futurs dans la presse en relation avec des attaques maritimes (Van Hage *et al.*, 2009), *etc.*

Une levée d'ambiguïté est réalisée en confrontant, en particulier, les lieux mentionnés dans le texte avec les lieux trouvés dans les métadonnées et dans les lieux connus du GEOLSIG. L'information ainsi extraite est chargée dans une base indexée par Elasticsearch qui permet de rechercher les entités géographiques semblant les plus pertinentes par rapport à celles repérées dans le texte. Une fois l'interprétation la plus proche choisie, le système retourne les coordonnées de l'entité géographique contenue dans le texte.

3.3.2. Amélioration de la reconnaissance des lieux

Nous avons vu dans l'introduction que dans le cadre du projet *Safecity* nous devons définir la localisation précise d'un événement. GEOLSemantics possède déjà un système de détection des lieux à base d'ontologie complété par des dictionnaires mais qui doit être amélioré. En effet, nos dictionnaires contiennent une liste importante de lieux mais ils ne peuvent pas être exhaustifs. Afin d'augmenter la qualité de la reconnaissance des lieux, nous utilisons des annonceurs de lieux tels que *rivière*, *rue*, *etc*. Cependant, certains lieux non introduits par un annonceur et dont le nom est trop ambigu ne sont pas reconnus par nos analyses. Par exemple, dans la phrase « le loup déborde », seul l'emploi du verbe *déborder* nous indique que *le loup* est un cours d'eau. Le principe est donc d'extraire du tweet une liste de lieux potentiels, en plus des lieux déjà détectés par l'analyse linguistique, puis d'interroger le GEOLSIG pour vérifier si ces lieux existent vraiment. Les nouveaux lieux trouvés sont ensuite réintroduits dans l'analyse afin de mieux reconnaître les événements géo-chrono-localisés.

En utilisant un ensemble de règles linguistiques, nous incluons plus de lieux potentiels, selon la présence ou non d'annonceurs de lieux, de majuscules ou de prépositions particulières. Une interrogation par *Elasticsearch* [6] pour chaque lieu potentiel va permettre de trancher si ces derniers sont effectivement des lieux ou non. Cette phase nous a permis de passer de 14% à 66% de lieux détectés.

Lors de cette interrogation, une requête peut retourner plusieurs solutions. Nous passons alors par une phase de désambiguïsation, en utilisant un système de poids

6. https://www.elastic.co/fr/elasticsearch

permettant de sélectionner le bon résultat parmi les solutions renvoyées. Le poids est la somme des éléments suivants : i) match complet ou partiel entre la requête et le résultat, et vice-versa, avec puis sans les mots vides, ii) catégorie du résultat dans la requête (ex : requête="Aéroport de nice", résultat "nice" catégorie="aéroport"), iii) catégorie présente dans la liste des catégories prioritaires ("ville, commune, quartier, avenue, boulevard, rue, place"), iv) diminution du poids si la catégorie est dans la liste des catégories minoritaires ("arrêt, bus, tram, café"). Le résultat avec le poids le plus important sera renvoyé. Si le poids est négatif aucun lieu ne sera renvoyé.

4. Chaîne de traitement GEOL

Le système d'émission d'alertes à partir des réseaux sociaux développé par GEOLSemantics dans le cadre du projet Safecity consiste en une analyse des tweets émis dans une zone géographique précisée par le client de l'application. Les modules vus précédemment se fondent sur une ontologie et des composants interconnectés présentés dans cette section.

4.1. GAL

Notre système se fonde sur une analyse linguistique profonde (*cf.* module GAL de la figure 1), réalisée grâce à un moteur développé en interne chez GEOLSemantics [7], afin de répondre aux exigences spécifiques à nos besoins. Nous analysons déjà le français, l'anglais et l'arabe. Bien évidemment, chaque nouvelle langue à traiter demande une adaptation des dictionnaires et des règles linguistiques d'extraction sémantique. Les tweets, quant à eux, sont rarement écrits dans un langage soigné mais remplis d'abréviations, sigles, argot, fautes d'orthographe, *etc*. Les premières étapes d'ana-

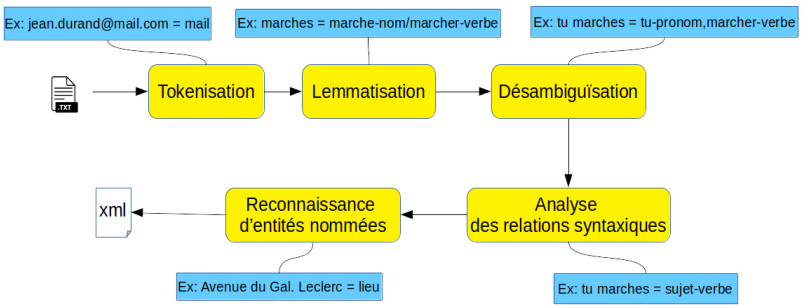


Figure 3. Processus de traitement de l'analyse linguistique profonde

lyse, la tokenisation, la lemmatisation et la désambiguïsation, sont traitées de façon classique mais en les adaptant aux spécificités des tweets.

7. http://www.geolsemantics.com/index.php/fr/

La **tokenisation** permet la reconnaissance des mots-dièse (dits *hashtags*) et des pseudonymes précédés d'une arobase. Cela simplifie l'analyse lorsque certains hashtags sont utilisés comme des mots appartenant à la phrase (par exemple, *"Une habitante se plaint de la présence des dépôts sauvages et des rats à #Courbevoie"*).

Beaucoup de travaux de recherche dans le contexte des tweets (Shoukry, Rafea, 2012) utilisent le *stemming*. Nous avons préféré passer par la **lemmatisation** des mots car nous possédons les ressources nécessaires (dictionnaires de formes fléchies) nous permettant des analyses plus précises. Le traitement des tweets a demandé des adaptations à plusieurs niveaux à savoir : (i) la mise en place d'un *dictionnaire spécialisé* contenant les mots argotiques, les abréviations et les sigles les plus courants, par l'analyse semi-automatique d'un corpus de tweets qui a permis d'intégrer 2179 entrées supplémentaires pour le français ; (ii) le *traitement grammatical contextualisé* des fins phonétiquement identiques : oubli du *s* du pluriel ou de la deuxième personne du singulier des verbes, confusion entre l'infinitif et le participe passé, etc. (iii) le remplacement de la désaccentuation par une *phonétisation* afin de rapprocher les mots lorsqu'ils contiennent des fautes d'orthographe. Par exemple, pour l'analyse de « Gros bazarre sur la #ligneR », après phonétisation le mot « bazarre » aura deux interprétations : *bazar* nom singulier, ou *bazars* nom pluriel ; (iv) le *découpage des mots collés de type hashtags*, en suivant la casse. Par exemple, dans un tweet tel que « La #TableRondeRATP est ouverte », le mot hashtag *#TableRondeRATP* permettra d'obtenir un découpage en *Table Ronde RATP*.

La **désambiguïsation** est effectuée grâce à une méthode statistique fondée sur des trigrammes et des bigrammes de catégories établis à partir d'un corpus étiqueté de tweets. Après la désambiguïsation, nous établissons les **relations syntaxiques** entre les différents éléments de la phrase (nom-adjectif, nom-complément de nom, mais aussi agent-action, action-objet, etc.), et nous détectons les **Entités Nommées.**

L'analyse linguistique nous a permis de lemmatiser et catégoriser les mots des tweets puis d'établir des relations entre les mots afin de simplifier l'étape suivante, c'est-à-dire l'extraction des événements et des informations qui leur sont propres. Aussi, notre système géolocalise les événements par le renforcement de la reconnaissance des entités nommées de lieux en faisant appel à un Système d'Information Géographique (*cf.* GEOLSIG dans la figure 1), créé à partir de fichiers de l'*open data* local et d'une extraction de fichier de l'*open data* global *Openstreetmap* [8].

4.2. PostGAL

L'analyse linguistique de GEOLSemantics (GAL) permet de collecter des noms de lieux. Cet outil repose sur un dictionnaire de noms de lieux connus. Si un nom de lieu est dans une phrase et n'est pas dans le dictionnaire, il est étiqueté en tant qu'Entité Nommée (EN) inconnue (*inc*). Une étape de désambiguïsation s'ensuit afin

8. https://www.openstreetmap.org/

d'attribuer une autre étiquette au mot inconnu (*i.e. loc, pers, org* pour respectivement lieu, personne ou organisation), en fonction de sa position dans la phrase et de ses relations avec les mots l'avoisinant. Par exemple, dans la phrase *Je vais à Kendira*, Kendira est absent du dictionnaire, mais étiqueté *EN loc* grâce au verbe aller. Bien que l'étape de désambiguïsation linguistique permette de collecter plus de lieux, elle reste très stricte en raison du nombre d'erreurs potentielles. Au final, il est préférable d'avoir plus d'*EN inc* que d'EN mal identifiées.

L'utilisation d'un outil de vérification de lieux améliore le taux de désambiguïsation. Avec ce nouveau système, nous interrogeons une base de données de lieux avec les *EN inc*. Si ce nom inconnu est présent dans la base, on peut le réinjecter dans le GAL sous une forme reconnaissable en tant qu'*EN loc*. Pour cela, nous avons établi une liste de règles linguistiques qui augmente la permissivité, et permet à plus de noms de lieux probables d'être collectés. Ces règles vont également permettre de compléter les noms de lieux et de potentiellement les regrouper avec des éléments les entourant. Par exemple, *Je suis au MK 2 de Bercy* devient *Je suis au MK 2 de Bercy* ; après complétion, *MK* est réuni avec *2*, puis *MK 2* avec *de Bercy*). En outre, beaucoup de lieux intègrent des noms de personnes. Là encore nous pouvons utiliser des règles linguistiques pour vérifier si une *EN pers* détectée n'est pas, en fait, un lieu. Il faut s'assurer que tous les composants du nom propre sont présents dans le lieu retenu. Par exemple, si l'on recherche *Nicolas Sarkozy*, il faut que *Nicolas* et *Sarkozy* soient présents dans le résultat.

La figure 4 présente l'architecture de ce système. Il s'agit de l'ensemble des traitements effectués après l'analyse linguistique, d'où le nom postGAL.

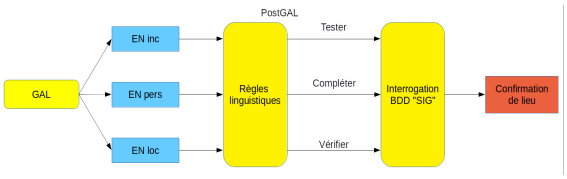


Figure 4. Organisation du traitement PostGAL

4.3. GEOLSIG

Comme mentionné en section 4.5, les événements découverts (grâce au module GEX) sont insérés dans une base et indexés par Elasticsearch. Ils y sont représentés au format JSON suivant une hiérarchie établie, à laquelle nous rajoutons plusieurs champs de dates. L'insertion dans Elasticsearch nous permet par la suite d'exécuter les requêtes pour la fusion d'événements, qu'il s'agisse de collecter, ajouter ou supprimer des entrées. Le contenu de la base est aussi utilisé pour la visualisation finale ou soumis à d'autres modifications. Un nettoyage de la BD est réalisé régulièrement dans le but de supprimer les événements (dits périmés) dont la date de validité est dépassée. Ainsi, la base conserve une taille réduite, assurant de meilleures performances pour les divers programmes l'utilisant. Par soucis d'anonymat, chaque événement est

caractérisé par un identifiant unique attribué avant son insertion dans la base et généré par nos soins.

4.4. GEX

Notre extraction (*cf.* module GEX dans la figure 1) permet de repérer des entités nommées (personnes, organisation, lieu, date, etc.) et des événements (incendie, accident, etc.) pouvant nécessiter l'intervention des autorités locales (pompiers, ambulance, police). Le GEX utilise des motifs syntactico-sémantiques guidés par une ontologie. Nous avons élaboré une ontologie permettant la modélisation des événements repérés dans les textes. Cette ontologie (*cf.* tableau 1) a été créée à l'aide de Protégé [9] au regard des spécifications du système d'alerte puis des différents corpus des événements. Cette ontologie fait partie d'une ontologie plus générale de la société GEOLSemantics, dans laquelle nous avons créé 170 classes, 363 ObjectProperties, 181 DatatypeProperties et déjà développé les définitions des entités nommées (de type lieu, date, *cf.* tableau 2.) relatives aux connaissances communes à divers domaines. [10].

Tableau 1. Ontologie partielle des événements

Concept	Sous-type d'action	Lieu	Date	Entité concernée
Incendie	event-type	event-place	event-date	isBurned
Accident	event-type	event-place	event-date	involve
Inondation	event-type	isFlooded	event-date	inundationSource
Colis suspect	package-type	event-place	event-date	–

Tableau 2. Ontologie partielle pour les entités

Concept	Propriétés
Location	loc-type, loc-name, locatedIn, hasGeoPosition
GeoPosition	coordinates
Date	dateBeg, dateEnd

Les patrons syntactico-sémantiques (dit pattern d'extraction) sont créés afin d'extraire les informations à partir des résultats de l'analyse linguistique profonde. Un pattern d'extraction est composé de trois éléments : i) un concept et une propriété définis par l'ontologie et permettant la représentation de l'information extraite au format RDF (Cyganiak *et al.*, 2014) ; ii) un patron lexico-syntaxique qui contient les informations morpho-syntaxiques d'une relation syntaxique ; iii) et un contexte constitué d'une ou plusieurs relations syntaxiques, aidant à la désambiguïsation de sens de l'action et de type de l'entité. Le GEX propose également une résolution des dates, consistant à transformer les dates relatives identifiées au niveau du module GAL comme *demain* ou *hier* en dates absolues en utilisant une date de référence. Il s'agit de la date absolue

9. voir https://protege.stanford.edu/

10. Cette ontologie n'est pas disponible pour des raisons commerciales

citée dans le texte ou, en cas d'absence de celle-ci, la date d'émission présente dans les métadonnées du tweet. Pour le cas particulier des tweets, lorsque les émetteurs du message ne précisent aucune date et que les verbes indiquent un événement en cours, alors nous attribuons à l'événement la date d'émission du tweet.

4.5. Event manager

Il est possible que des événements extraits soient similaires. Pour cette raison, chaque événement nouvellement extrait doit être comparé avec ceux précédemment collectés. S'ils s'avèrent similaires, ils sont alors fusionnés. La fusion est effectuée lorsqu'il y a une intersection temporelle et spatiale entre les événements, et que les actions qu'ils représentent sont identiques. Afin de vérifier une telle similarité, nous procédons en deux temps. Une requête permet de récupérer dans notre base de données les événements ayant des actions identiques et dont les intervalles de validité se chevauchent, puis l'intersection spatiale est calculée en utilisant une bibliothèque Java appropriée. Nous aurions également pu ajouter une clause à notre requête pour vérifier ce dernier point, mais le système d'indexation Elasticsearch [11] que nous utilisons ne permet pas de calculer l'intersection entre les différents types de surfaces, chose nécessaire pour la représentation fusionnée.

Lorsqu'ils sont regroupés, les événements conservent la liste des tweets qu'ils représentent. Les dates et les lieux doivent alors être précisés: pour les dates de début et de fin d'événement, nous considérons les valeurs les plus anciennes comme étant les plus précises. En ce qui concerne les lieux, l'intersection des représentations est adoptée, et dans le cas d'événements localisés par des points, nous effectuons une vérification de distance à la place de l'intersection, en utilisant une valeur de référence paramétrable. Nous considérons qu'un point est toujours jugé plus précis par rapport à un polygone ou un multipoint. Pour la comparaison d'événements représentés chacun par un point, si la distance les séparant est inférieure à la valeur de référence, nous considérons qu'il y a intersection, et utilisons le centre du segment formé par les deux points comme localisation dans la fusion des événements.

Lors de la fusion, le reste des métadonnées de l'événement généré provient de l'événement le plus ancien. Néanmoins, puisque chaque événement contient la liste des tweets qui l'a généré, l'ensemble des informations de base reste disponible dans le résultat final pour tout utilisateur. Nous avons appelé le composant responsable de ce traitement l'*Event Manager* (*cf.* figure 1).

11. voir https://www.elastic.co/fr/Elasticsearch

5. Évaluations

5.1. Expérimentations

BERT [12] (Devlin *et al*., 2019) est un modèle contextuel de langues, fondé sur les réseaux de neurones, développé chez Google et diffusé en *open source* en 2018. Le succès révolutionnaire de BERT et de la classe de modèles appelés "transformeurs" (BERT, ALBERT, RoBERTa, GPT, GPT-2, etc.), a permis de mettre au point des modèles linguistiques capables d'atteindre des performances record dans les différentes tâches du NLU "Natural Language Understanding". CamemBERT et FlauBERT sont deux variantes qui offrent les meilleurs résultats pour la langue française. L'avantage d'utiliser ces modèles est la réduction du coût de préparation du corpus d'apprentissage, car ils sont pré-entraînés sur des corpus très volumineux en français issus de différentes sources (Wikipedia, livres, internet, journaux, etc.). Notons que le plongement contextuel (*Contextual embeddings*), comme BERT, va au-delà des représentations des plongements de mots (*Word embedding*), comme Word2Vec, et atteint des performances révolutionnaires sur un large éventail de tâches de traitement du langage naturel puisqu'il attribue à chaque mot une représentation basée sur son contexte. (Liu *et al*., 2020) ont examiné les modèles existants d'intégration contextuelle.

Pour la comparaison avec notre système, nous avons affiné le réseau de neurones fourni par CamemBERT en lui ajoutant une couche de sortie concernant la reconnaissance des deux entités "ACTION" et "LOCATION". Nous avons entraîné ce nouveau modèle sur un corpus d'apprentissage composé de 1195 tweets sur les inondations dans la région de la Côte d'Azur en France, sur une période allant de 2012 jusqu'à 2019. Le corpus intégral est composé de 1495 tweets, annotés en précisant le lieu, le déclencheur de l'action *inondation* (inondation, montée des eaux, etc.), et si le tweet représente un événement ou non. 300 tweets ont été isolés après annotation et utilisés comme corpus de test pour les deux systèmes.

Pour notre solution, nous avons pris un corpus de développement de 118 tweets (10%) issus du corpus d'apprentissage de BERT pour affiner nos règles. Après l'affinage des règles sur les deux systèmes à comparer, nous avons lancé les analyses sur le corpus de test restant de 300 tweets. Les résultats sont présentés dans le tableau 3.

Afin de valider les fusions d'événements, plusieurs des tweets utilisés présentaient des similarités partielles ou totales (considérant l'action et les localisations temporelles et géographiques). Ceci nous a permis de confirmer le bon fonctionnement de la fusion de tweets. Les résultats ont également pu être analysés afin de s'assurer que l'intégralité des champs JSON du regroupement - identifiant, dates, intersection - étaient bien remplis et au bon format.

12. Bidirectional Encoder Representations from Transformers

5.2. Analyse des résultats

Dans un premier temps, nous avons remarqué que BERT est efficace pour détecter des lieux et des actions mais présente des lacunes en sémantique. La F-mesure de la détection des actions seules (resp. lieux) est de 88% contre 70% pour GEOLSemantic (resp. 96% contre 88%). En effet, un message peut inclure un déclencheur d'action et un lieu sans pour autant représenter un événement (négation, événement très ancien, métaphore, etc.), tandis que notre système a une plus grande précision (88%) et ne renvoie que les événements sémantiquement corrects. Par exemple, *"Mon balcon est inondé, c'est une piscine olympique le truc"* est renvoyé par BERT, mais pas par GEOLSemantics.

Puisque notre outil est plus correct dans son traitement des événements qui ne doivent pas déclencher d'alerte, nous avons pris les résultats de GEOLSemantics (GEOL) sur les non-événements et les résultats de BERT sur les événements, constituant ainsi un système hybride qui arrive à 89% de F-mesure tout en étant meilleur en rappel et en précision que chacun des systèmes exécutés individuellement. Enfin nous avons étudié les cas où au moins l'un des systèmes avait une bonne réponse. Cette approche a obtenu 95% de F-mesure. Il semble que les deux systèmes soient complémentaires et qu'il faille les configurer selon le contexte. Concernant les temps d'exécution, BERT est plus rapide, ce qui s'explique par la quantité de traitements que nous effectuons. Nous pallions ceci en effectuant de la parallélisation des traitements à tous les niveaux d'analyse, mais aussi en ajoutant une première étape qui applique un filtre « à grandes mailles » sur le flux des tweets d'entrée, afin d'analyser que ceux qui ont une chance de contenir des informations pertinentes. Ce filtre s'appuie sur une liste de mots-clés pouvant représenter un événement à extraire.

Tableau 3. Tests de performance des systèmes GEOLSemantics et BERT

	GEOL	Bert	Non événement–>GEOL évenement–>BERT	Combinaison GEOL/BERT
Rappel	86%	90%	90%	99%
Présision	88%	74%	89%	91%
F-mesure	87%	80%	89%	95%

6. Conclusion et perspectives

Nous avons montré qu'un processus de traitement complet combinant des approches linguistique et sémantique afin de compléter les informations composant un événement par recoupement et regroupement est possible. Un système hybride nous permet d'obtenir des résultats très satisfaisants pour une première expérience, résultats améliorables en configurant mieux BERT et en rendant notre système encore plus robuste.

Diverses améliorations sont planifiées pour l'algorithme de regroupement d'événements ayant un contexte commun partiel. Par exemple, regrouper des événements de

localisation géographique et de temporalité similaires mais dont l'action représentée est différente. En effet, malgré cette divergence d'action, les autres éléments détectés laissent suggérer que de tels événements sont liés d'une manière ou d'une autre. D'autre part, il faut affiner la modélisation des zones géographiques des événements à fusionner. En effet, dans plusieurs cas, utiliser l'intersection entre chacun des lieux déterminés s'avère moins pertinent que de conserver une zone trouvée dans un des événements à fusionner.

Dans cet article, nous avons expérimenté notre système sur un corpus d'inondation. Pour nos travaux futurs, nous souhaitons étudier la pertinence de notre système dans d'autres contextes. En effet, le tableau 3 a montré que d'autres contextes peuvent influencer le choix du système. Un nouveau corpus (7627 tweets sur des incendies sur la Côte d'Azur pour une période allant de 2012 à 2020) va également nous permettre d'étudier les capacités de BERT pour la détection des multi-événements, tâche déjà développée dans notre système. Nous allons aussi tester la rapidité et la précision des deux systèmes en simulant un flux de tweets en temps réel comportant beaucoup de bruit (tweets non pertinents).

Remerciements

Cet article est dans le cadre du projet Safecity soutenu par la BPI. Nous remercions Houda Saadane, El Mehdi Khalfi et Christian Fluhr pour leur implication.

Bibliographie

Aramaki E., Maskawa S., Morita M. (2011). Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, p. 1568–1576.

Besnard P., Cordier M.-O., Moinard Y. (2008). Ontology-based inference for causal explanation. *Integrated Computer-Aided Engineering*, vol. 15, n° 4, p. 351–367.

Broniatowski D. A., Paul M. J., Dredze M. (2013). National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PloS one*, vol. 8, n° 12.

Cyganiak R., Wood D., Lanthaler M. (Eds.). (2014). *Rdf 1.1 concepts and abstract syntax.* Consulté sur https://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

Devlin J., Chang M.-W., Kenton L., Toutanova K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In . Consulté sur https://arxiv.org/pdf/1810.04805.pdf

Guibon G., Ermakova L., Seffih H., Firsov A., Le Noé-Bienvenu G. (2019). Multilingual fake news detection with satire.

Han B., Cook P., Baldwin T. (2013). A stacking-based approach to twitter user geolocation prediction. In *Proceedings of the 51st annual meeting of the association for computational linguistics: system demonstrations*, p. 7–12.

Hecht B., Hong L., Suh B., Chi E. H. (2011). Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proceedings of the sigchi conference on human factors in computing systems*, p. 237–246.

Hoang T. B. N., Mothe J. (2018). Location extraction from tweets. *Information Processing & Management*, vol. 54, n° 2, p. 129–144.

Huang B., Carley K. M. (2019). A hierarchical location prediction neural network for twitter user geolocation. *arXiv preprint arXiv:1910.12941*.

Jackoway A., Samet H., Sankaranarayanan J. (2011). Identification of live news events using twitter. In *Proceedings of the 3rd acm sigspatial international workshop on location-based social networks*, p. 25–32. New York, NY, USA, Association for Computing Machinery. Consulté sur https://doi.org/10.1145/2063212.2063224

Khamis S., Ang L., Welling R. (2017). Self-branding,'micro-celebrity'and the rise of social media influencers. *Celebrity studies*, vol. 8, n° 2, p. 191–208.

Kleinberg J. (2003). Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, vol. 7, n° 4, p. 373–397.

Kryvasheyeu Y., Chen H., Moro E., Van Hentenryck P., Cebrian M. (2015). Performance of social network sensors during hurricane sandy. *PLoS one*, vol. 10, n° 2.

Liu Q., Kusner M. J., Blunsom P. (2020). A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*.

Perez S. (2017, Nov). *Twitter officially expands its character count to 280 starting today.* TechCrunch. Consulté sur https://techcrunch.com/2017/11/07/twitter-officially-expands-its-character-count-to-280-starting-today/?guccounter=1

Sakaki T., Okazaki M., Matsuo Y. (2010). Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on world wide web*, p. 851–860.

Shoukry A., Rafea A. (2012). Preprocessing egyptian dialect tweets for sentiment mining. In *The fourth workshop on computational approaches to arabic script-based languages*, p. 47.

Steinert-Threlkeld Z. C. (2018). *Twitter as data*. Cambridge University Press. Consulté sur https://www.cambridge.org/core/elements/twitter-as-data/27B3DE20C22E12E162BFB173C5EB2592

Troncy R., Shaw R., Hardman L. (2010). Lode: une ontologie pour représenter des événements dans le web de données. In *21es journées francophones d'ingénierie des connaissances (ic 2010), http://www. ic2010. mines-ales. fr/*, p. 69–80.

Van Hage W. R., Malaisé V., Vries G. de, Schreiber G., Someren M. van. (2009). Combining ship trajectories and semantics with the simple event model (sem). In *Proceedings of the 1st acm international workshop on events in multimedia*, p. 73–80.

Wang X., Gerber M. S., Brown D. E. (2012). Automatic crime prediction using events extracted from twitter posts. In *International conference on social computing, behavioral-cultural modeling, and prediction*, p. 231–238.

Weng J., Lee B. (2011, 01). Event detection in twitter.

Ying Y., Peng C., Dong C., Li Y., Feng Y. (2018). Inferring event geolocation based on twitter. In *Proceedings of the 10th international conference on internet multimedia computing and service.* New York, NY, USA, Association for Computing Machinery. Consulté sur https://doi.org/10.1145/3240876.3240909

Analyse des discours sur Twitter dans une situation de crise

Étude de l'incident à l'usine Lubrizol de Rouen

Hiba Abou Jamra, Annabelle Gillet, Marinette Savonnet, Éric Leclercq

Laboratoire d'Informatique de Bourgogne - EA 7534
Univ. Bourgogne Franche-Comté
9, Avenue Alain Savary, F-21078 Dijon - France
Hiba_Abou-Jamra@etu.u-bourgogne.fr

RÉSUMÉ. *Les données des réseaux sociaux ont un potentiel de valeur que les outils d'analyse doivent révéler. Twitter, en tant que plateforme de microblogging, facilite les interactions entre ses utilisateurs, permet la diffusion rapide d'informations. L'analyse des grandes tendances, des évènements, à partir des données permet de comprendre ou d'agir sur le monde réel. Dans cet article, nous présentons une méthodologie de travail et des résultats d'analyse pour l'étude de la communication dans les domaines de l'alimentation et de la santé qui ont pour cadre le projet inter-disciplinaire COCKTAIL incluant plusieurs partenaires industriels. Nous identifions un cas d'utilisation général pour différents types d'analyses des discours et nous montrons comment les algorithmes permettent de construire un observatoire afin de proposer des indicateurs macroscopiques et d'étudier les évènements, les communautés, les utilisateurs influents. Un exemple concret autour de l'évènement Lubrizol servira de fil conducteur pour illustrer les fonctionnalités et les problématiques traitées.*

ABSTRACT. *Social media data has value potential that analytics tools need to reveal. Twitter, as a microblogging platform, facilitates interactions between its users, allows rapid dissemination of information. The analysis of major trends, events, from the data makes it possible to understand or act on the real world. In this article, we present a working methodology and analysis results for the study of communication in the fields of food and health, which are part of the interdisciplinary COCKTAIL project involving several industrial partners. We identify use cases for different types of analysis and we show how the analysis algorithms make it possible to build an observatory in order to propose macroscopic indicators and to study events, communities, influential users. A concrete example around the Lubrizol event will serve as a common thread to illustrate the functionalities and the addressed problems.*

MOTS-CLÉS : *Analyses des réseaux sociaux, Twitter, détection des communautés, mesures de centralité, détection des évènements, séries temporelles*

KEYWORDS: *Social network analysis, Twitter, community detection, centrality measure, event detection, time series*

1. Introduction et problématique

Les données des réseaux sociaux sont créées à partir des interactions entre les individus, ces interactions étant amplifiées par les relations personnelles. Ces réseaux possèdent des caractéristiques intéressantes en terme de valorisation des données. Cependant, ils ont des propriétés spécifiques (distribution en loi de puissance, petit monde, assortativité, attachement préférentiel, etc.) qui nécessitent des outils d'analyse plus sophistiqués que les approches classiques des systèmes d'information ne proposent pas. Par exemple, la structure communautaire des réseaux sociaux est une des propriétés fondamentales. Cette structure peut être utilisée pour comprendre les interactions entre les utilisateurs mais aussi pour expliquer des évènements. En effet, des observations ont mis en évidence que certains évènements émergent plus vite à travers les réseaux sociaux qu'à travers d'autres médias plus traditionnels comme les sites Web, la radio et la télévision (Aiello *et al*., 2013). Dans ce cadre, Twitter est reconnu comme un révélateur d'évènements importants souvent quelques minutes ou heures après qu'ils se soient produits (Fedoryszak *et al*., 2019) et comme une chambre de résonance à forte influence qui propage rapidement l'information dans des communautés polarisées.

Notre objectif est de proposer une plateforme de collecte et d'analyse prenant en compte la richesse des données Twitter afin d'appréhender les usages, la circulation de l'information et la construction des discours. Si le fonctionnement de Twitter est simple – des tweets produits par des utilisateurs à l'aide de quelques opérateurs – les liens créés sont, selon le domaine ou les intentions de l'utilisateur, sémantiquement différents (Azaza *et al*., 2019). Les workflows doivent être reproductibles, pour cela nous utilisons Jupyter [1] pour construire des squelettes d'analyses dans lesquels chaque série de traitements est clairement identifiée : provenance des données, description des données de sortie, algorithmes, conditions d'application des algorithmes, contraintes pour l'interprétation des résultats produits.

Notre démarche de travail repose sur différents niveaux d'analyse (macroscopique, mésoscopique puis microscopique) et cherche à comprendre comment les espaces des utilisateurs et des hashtags sont constitués et comment ils interagissent, ceci afin d'éclairer la structure de la communication sur Twitter. Elle sera décrite à partir de l'étude d'un évènement qui est l'incendie de l'usine Lubrizol à Rouen. Pour l'analyse macroscopique, des chiffres significatifs et des fréquences brutes seront identifiés, puis l'étude mésoscopique se concentrera sur les communautés d'utilisateurs et la centralité de ces derniers dans chacune des communautés et enfin pour l'analyse microscopique nous analyserons le cas particulier de la construction du discours de responsabilité institutionnelle lors d'une crise.

1. https://jupyter.org

L'article est organisé comme suit, la section 2 présente des travaux similaires traitant de l'analyse d'évènements et de crises sur Twitter. La section 3 présente les principes du projet interdisciplinaire COCKTAIL, nous y discutons de la collecte et du nettoyage des données. La section 4 détaille notre approche appliquée aux tweets concernant l'incendie de l'usine Lubrizol à Rouen avec des analyses sur l'utilisation des hashtags, de la structuration des communautés d'utilisateurs et comment les institutions (État, métropole, agence sanitaire et de santé, rectorat, etc.) ont réagi et communiqué. La section 5 présente la mise en œuvre des workflows reproductibles à l'aide de l'outil Jupyter. Finalement, la section 6 conclut l'article et dresse les perspectives dégagées par ce travail.

2. Travaux connexes

Twitter est l'un des réseaux sociaux les plus utilisés, faisant de lui un modèle représentatif pour les Sciences Humaines et Sociales. Plusieurs études ont analysé les discours sur Twitter afin d'identifier les réactions des individus face à des crises naturelles ou des évènements remarquables dans la politique, le sport, l'économie ou la société. Nous pouvons citer les études de MacEachren *et al*. (2011) qui analysent dans le cadre d'une crise les activités dans l'espace et le temps des utilisateurs de Twitter, et Öztürk et Ayvaz (2018) qui ont exploré les opinions et les sentiments des utilisateurs de Twitter à l'égard de la crise des réfugiés syriens. Dans la suite de cette section, nous présentons plus en détail des travaux sur le Brexit qui se rapprochent de notre démarche.

Le Brexit a donné lieu à plusieurs études ayant pour but d'analyser ses causes et ses conséquences à partir des messages échangés sur Twitter. Mora-Cantallops *et al*. (2019) ont étudié l'influence du Brexit sur la façon dont les informations sont discutées sur le réseau, mais également la création de messages et la forme du réseau lui-même. 4 037 684 tweets, émis par des utilisateurs localisés au Royaume-Uni et catégorisés en positif (quitter), négatif (rester) ou neutre, ont été collectés entre le 12 Mai et le 23 Juin 2016. Ces tweets, regroupés par intervalles d'une heure, ont permis de construire des graphes dont les nœuds sont les utilisateurs et les liens sont les relations retweet, reply et quotes. Ensuite, pour comprendre la volatilité des interactions face à un tel évènement, les auteurs ont utilisé le modèle statistique GARCH (*Generalized Autoregressive conditional heteroscedasticity*) pour analyser les séries temporelles. Des mesures de centralité pour détecter les utilisateurs influents et des détections de communautés ont aussi été effectuées. La limite de cette étude est que les données sont incomplètes car les tweets étudiés ont été obtenus à partir d'un processus d'échantillonnage.

Vasiliu *et al*. (2016) ont surveillé et analysé les interactions sur Twitter autour du Brexit en utilisant la plateforme du projet SSIX (*Social Sentiment analysis financial IndeXes*), qui fournit aux entreprises européennes un ensemble de logiciels afin d'analyser et de comprendre les sentiments exprimés sur les médias sociaux. Les données ont été collectées à partir de 75 critères (mots-clés, hashtags et utilisateurs) entre le 4

et 30 Mai 2016, et archivées dans une base non-relationnelle. L'analyse s'est déroulée sur trois périodes: 1) avant le vote, 2) le jour du vote, et 3) après le vote. Pour les périodes avant et après le vote, les auteurs ont observé les tendances sur une période de 3 à 4 jours, alors que pour le jour du vote, ils observaient les tendances toutes les deux à trois heures. Les tweets ont été classés en deux catégories : quitter et rester. Ils ont découvert que leur résultat différait du vote de 9,4%, probablement en raison de la tranche d'âge des utilisateurs de Twitter, de la localisation des tweets collectés et du fossé éducatif entre les utilisateurs de Twitter et la population britannique.

Nous remarquons que les travaux décrits précédemment ont étudié les données de Twitter, pour un évènement donné en utilisant un ensemble d'outils spécifiques pour l'analyse. Il est important de noter que ces études se sont concentrées sur des critères particuliers du réseau Twitter comme par exemple la polarisation, ce qui les rend incomplètes et peu reproductibles en termes d'analyses des discours relatifs à des évènements et des incidents importants.

3. Le projet COCKTAIL : contexte, objectifs

Le projet COCKTAIL, lauréat d'un appel à projet ISITE-BFC (*Initiatives Science Innovation Territoires Économie en Bourgogne-Franche-Comté*), vise à créer un observatoire en temps réel des tendances, des singularités et des signaux faibles circulant dans les discours sur Twitter. Les analyses effectuées ont pour objectif de détecter les discours critiques pouvant devenir viraux, les communautés clés d'acteurs liées au domaine étudié (alimentation et santé) avec leurs liens/interactions, les évènements ou précurseurs d'évènements, les discours émergents caractérisés par des signaux faibles ainsi que les tendances culturelles liées aux pratiques alimentaires et à la santé. Le consortium comprend des chercheurs en Sciences de l'Information et de la Communication, en Informatique, en Sciences Cognitives, et en Sciences des Aliments, le pôle de compétitivité Vitagora spécialisé dans l'agro-alimentaire et des entreprises du domaine informatique. L'observatoire, logiciel libre, permettra aux chercheurs et aux stratèges des secteurs public et privé de bénéficier des avancées scientifiques du projet et aux partenaires de développer des services adaptés à leur modèle économique.

Le diagramme de cas d'utilisation (*use case*) de la figure 1 présente les acteurs et les grandes fonctionnalités de la plateforme. COCKTAIL fait intervenir plusieurs acteurs dont le *data engineer* qui conçoit et gère l'architecture de collecte et de stockage des données. Il classe les données recueillies en fonction des besoins exprimés par le commanditaire et aide le *data scientist* dans la phase de nettoyage des données. Le *data analyst* et le *data scientist* traduisent et formalisent le questionnement métier du commanditaire. Le *data scientist* croise les données, sélectionne, modifie les algorithmes et valide les analyses. Finalement le *data analyst* transforme les analyses en informations métier pour le commanditaire.

Dans la suite, nous expliquons brièvement l'architecture Hydre développée pour la collecte en continu et le stockage des données issues de Twitter (Gillet *et al.*, 2019). Un système de stockage de type polystore a été développé et intégré dans la plateforme.

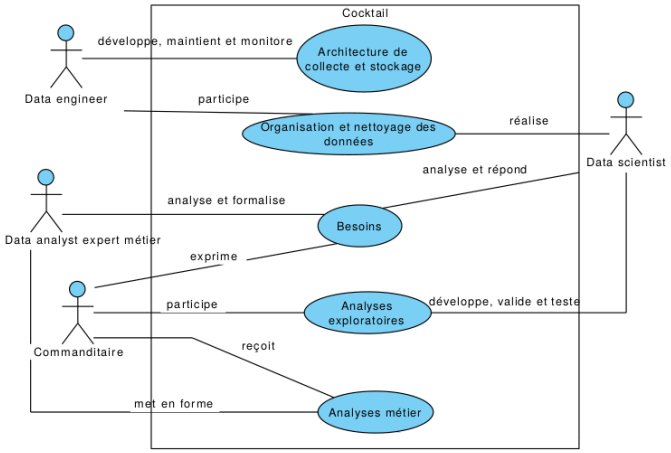


Figure 1. Diagramme de cas d'utilisation de niveau 1 décrivant les fonctionnalités de la plateforme COCKTAIL et les acteurs correspondants

L'architecture de la plateforme s'inspire du patron de Lambda Architecture (Marz, Warren, 2015) et s'appuie sur les composants suivants :

– des processus de collecte implémentés avec le modèle d'acteurs Akka [2] et déployés sur un cluster. Ces processus exploitent les deux APIs que Twitter met à disposition : l'API search et l'API stream. Ces APIs sont exploitées avec la librairie Twitter4j. Les critères de collecte sont des comptes, des hashtags et des mots-clés. Le système Kafka [3] est utilisé pour servir d'intermédiaire entre la partie collecte et les couches Speed et Batch et permet de conserver les données pendant 7 jours, ce qui contribue à la capacité de résistance aux pannes de l'architecture ;

– la couche Speed permet de calculer des indicateurs macroscopiques en temps réel en utilisant Kafka Streams [4], afin d'établir des séries temporelles en temps réel sur les éléments importants constitutifs des tweets comme les hashtags, les mentions et de servir de base pour détecter des évènements ;

– la couche Batch enregistre les données brutes dans Hadoop HDFS (*Hadoop Distributed File System*) [5] et parallèlement par micro-batch dans le polystore. Le stockage dans HDFS permet de lancer des traitements de reprises depuis les données brutes si le schéma de données du polystore est modifié suite à des évolutions fonctionnelles ou techniques ;

– la couche Serving est implantée par un polystore qui met à disposition les données dans des formats adaptés à leur analyse. Le polystore comprend les bases de don-

2. https://akka.io/
3. https://kafka.apache.org/
4. https://kafka.apache.org/documentation/streams/
5. https://hadoop.apache.org/

nées PostgreSQL [6] (pour le traitement des données attributaires), ArangoDB [7] (pour l'analyse des graphes) et TimescaleDB [8] (pour l'analyse des séries temporelles).

L'architecture Hydre est capable de supporter le flux moyen de Twitter (6 000 tweets par seconde) permettant des collectes importantes. À la fin février 2020, le nombre de tweets collectés dans le polystore atteint environ 75M, pour un volume de 1,9 To de données brutes JSON et 171 Go dans PostgreSQL.

Hydre est la suite de SNFreezer, une plateforme multi-paradigme de collecte, de stockage et d'analyse de tweets (Basaille *et al*., 2016), dans laquelle nous avons comparé les différentes fonctionnalités de SNFreezer avec certains projets concurrents.

4. Étude de cas : l'incendie de l'usine Lubrizol à Rouen

Dans cette section, nous étudions les prises de forme des discours sur Twitter dans le cadre de l'incendie de l'usine Lubrizol à Rouen, et nous analysons plus particulièrement la construction du discours de responsabilité institutionnelle suite à cet incendie. Pour cette étude nous développons une méthodologie d'analyse sur trois niveaux (macroscopique, mésoscopique et microscopique), qui est implantée sous la forme de *notebooks* dans Jupyter.

4.1. Chronologie de l'incident et première apparition sur Twitter

Le 26 septembre 2019, entre 2h40 et 2h50 du matin, une partie de l'usine Lubrizol à Rouen et trois bâtiments de Normandie Logistique ont été ravagés par un incendie qui a provoqué une énorme fumée noire de 22 km. Lubrizol [9], fabriquant des additifs pour lubrifiants industriels et pour les carburants, est classée Seveso « à haut risque ». Le sinistre a fait aucune victime, mais sa cause reste pour le moment inconnue, même la localisation précise du départ du feu n'a pas pu être déterminée [10]. Dès le début, Twitter a joué un rôle très important dans la diffusion de l'information sur cet incendie. Alors que l'incendie a débuté vers 2h40, le premier tweet relatif à l'incendie a été publié à 3h03 par Thomas Schonheere, journaliste à France Bleu Normandie, le mot lubrizol apparaît à 3h24, et à 4h15 le hashtag `#lubrizol` surgit pour la première fois, le premier tweet d'une institution (le préfet de la Seine Maritime) est émis à 4h50 demandant d'éviter le secteur.

6. https://www.postgresql.org/
7. https://www.arangodb.com/
8. https://www.timescale.com/
9. https://france.lubrizol.com/fr-FR/About
10. https://www.lexpress.fr/actualite/societe/incendie-de-l-usine-lubrizol-a-rouen-ce-que-l-on-sait-un-mois-apres_2104941.html

4.2. Nettoyage des données

L'intérêt de l'évènement nous a amené à lancer une collecte pour laquelle 47 hashtags et 111 utilisateurs ou comptes ont été fournis par les chercheurs en sciences humaines et sociales du projet COCKTAIL comme critères de collecte :

Hashtags : `#CELubrizol`, `#lubrizolrouen`, `#lubrizoltransparence`, `#seveso`, `#VeritePourRouen`, `#WarrenBuffettDoitPayer`, etc.

Utilisateurs : `@atmonormandie`, `@fbleunormandie`, `@Min_Ecologie`, `@prefet76`, `@regionNormandie`, `@damienadam76`, etc.

Ces critères ont permis de collecter environ 2 millions de tweets entre le 26 septembre et le 26 novembre 2019. Pour diminuer le bruit dans la collecte, nous avons nettoyé ces données en filtrant celles nécessaires pour notre étude. Dans cette étape, nous avons gardé les tweets correspondants aux critères fournis (hashtags et comptes) filtrés avec les mots `lubrizol` ou `rouen`. Par exemple nous retenons les tweets contenant `lubrizol` ou `rouen` avec `controlessanitaires`, ou encore `rouen` avec `desastreecologique`. En revanche avec cette méthode les tweets du compte `@fbleunormanandie` sans rapport avec l'incident sont éliminés. À l'issue de cette opération, le corpus est réduit à 558 895 tweets. Parmi ces tweets, 73 126 sont des tweets originaux et 485 769 sont des retweets émis par 141 177 comptes. Durant la seule journée du 26 septembre, 29 495 tweets dont 26 940 retweets ont été produits sur l'évènement, confirmant bien l'effet chambre d'écho de Twitter.

4.3. Exploration macroscopique du corpus

Tableau 1. TOP 5 des utilisateurs institutionnels les plus actifs

Utilisateurs	Nombre de tweets originaux
Sénat Direct	61
Sénat	58
Préfet de la Seine-Maritime	53
Atmo Normandie	20
Gouvernement	17

Utilisateurs	Nombre de retweets
Métropole Rouen Ndie	78
Ministère de la Solidarité et de la Santé	41
Ville de Rouen	35
Préfet de la Seine-Maritime	28
ARS Normandie	27

Utilisateurs	Nombre de tweets retweetés
Préfet de la Seine-Maritime	135
Gouvernement	43
Minitère de l'intérieur	25
ARS Normandie	15
Académie de Rouen	8

Nous avons réalisé des analyses exploratoires de niveau macroscopique sur le corpus en calculant des chiffres significatifs comme le nombre total des tweets comprenant des tweets originaux, des retweets, des réponses et des citations. Ainsi, nous avons calculé les fréquences d'apparition des producteurs de tweets, des retweeteurs, des utilisateurs qui sont les plus retweetés et des hashtags, à l'aide de requêtes analytiques SQL lancées sur la base de données relationnelle. Le tableau 1 montre les utilisateurs institutionnels les plus actifs. Afin d'étudier comment ces utilisateurs communiquent, nous avons construit le graphe d'interaction $utilisateur-utilisateur_{retweet}$, et nous avons ensuite généré une visualisation avec le logiciel Gephi [11]. Sur la figure

11. https://gephi.org/

2, la taille des nœuds est proportionnelle à leur centralité. Il est à noter que les utilisateurs `Senat` et `Senat_direct` sont isolés, cela peut s'expliquer par la constitution d'une commission d'enquête sénatoriale sur Lubrizol le 10 octobre.

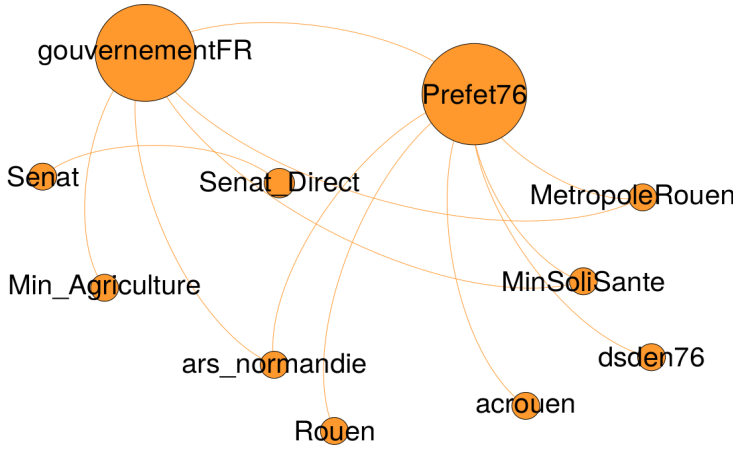


Figure 2. Graphe de la relation retweet pour les utilisateurs institutionnels les plus actifs

Nous avons réalisé une étude globale macroscopique des hashtags.La forte présence du hashtag `#lubrizol` nous a amené à travailler sur les co-occurences de celui-ci ainsi que ses composés. Le tableau 2 montre l'inquiétude concernant les risques encourus avec l'apparition des hashtags `ecologie`, `hydrocarbures` et `climat`. Cette inquiétude est associée à un besoin de transparence mais aussi à une rapide politisation d'après les hashtags `climatstrike`, `giletsjaunes`, et `act46` et à l'interpellation du pouvoir en place avec les hashtags `gouvernement`, `castaner`, `larem` et `enmarche`. Il est à remarquer un détournement de ce hashtag pour mettre en avant l'acte 46 des gilets jaunes (manifestation du 28 septembre). En effet, 14 834 utilisateurs différents ont relayé ou émis des tweets relatifs aux gilets jaunes dans notre corpus.

4.4. Analyses mésoscopiques

Dans ce qui suit, nous allons décrire la détection des communautés dans différents types de graphes, le calcul des scores de hubs et autorités, et les mesures de centralité pour étudier les phénomènes d'influence.

4.4.1. Détection des communautés

Un processus pertinent dans l'analyse des réseaux sociaux est de découvrir des nœuds qui partagent des caractéristiques communes dans un graphe. Ce processus, appelé détection de communautés, consiste à trouver des groupes ayant une forte densité de liens. Plusieurs travaux ont proposé différentes méthodes de détection de com-

Tableau 2. Trois co-occurents du hashtag `#lubrizol` *et ses composés*

#lubrizol et composés	Hashtag2	Hashtag3	Hashtag4	Fréquence
lubrizol	incendierouen	giletsjaunes	act46	2363
lubrizol	incendierouen	giletsjaunes	acheres	1303
lubrizolrouen	lubrizol	gouvernement	castaner	905
lubrizol	ecologie	climat	castaner	904
lubrizol	incendie	hydrocarbures	actionclimat	760
lubrizoltransparence	lubrizolrouen	lubrizol	gouvernement	757
lubrizol	ecologie	climatestrike	climat	750
lubrizol	lubrisol	giletsjaunes	act46	655
lubrizoltransparence	lubrizolrouen	lubrizol	larem	592
lubrizol	lrem	enmarche	agriculteurs	515

munautés. Nous avons utilisé l'algorithme Louvain basé sur la maximisation de la modularité (Blondel *et al.*, 2008). Nous avons choisi cet algorithme car il s'exécute en O(nlogn) et il exploite des informations globales sur la topologie du réseau. Nous avons aussi employé l'algorithme Walktrap créé par Pons et Latapy (2005) pour détecter des communautés. Il se base sur l'idée qu'une marche aléatoire à partir d'un nœud reste pendant un certain temps dans la communauté du nœud. Nous avons utilisé deux algorithmes comparables car ils utilisent la modularité mais s'appuient sur des principes différents. Le tableau 3 synthétise les résultats produits par les deux méthodes pour le graphe $utilisateur - utilisateur_{retweet}$. Nous remarquons que les méthodes Louvain et Walktrap trouvent deux communautés identiques qui correspondent pour l'une aux médias nationaux comme `BFMTV`, `FranceInfo` et `CNEWS` et pour l'autre aux médias locaux comme `76actu` et `paris_normandie`. Les tailles des communautés correspondantes appuient cette ressemblance. Les nœuds `AiphanMarcel`, `Loran076` , `CerveauxNon` et `Rouendanslarue` appartenant à une communauté détectée par Walktrap se retrouvent dans deux communautés pour Louvain. L'utilisateur `DuPouvoirDachat` est fortement considéré central dans l'une des communautés avec Louvain, tandis que nous ne le retrouvons pas comme un nœud central (importance négligeable) avec Walktrap (valeur `NA`).

Tableau 3. Détection de communautés avec Louvain et Walktrap dans le graphe $utilisateur - utilisateur_{retweet}$

Utilisateur	Communauté Louvain	Communauté Walktrap
AiphanMarcel	Taille communauté: 550	Taille communauté: 716
Loran076		
CerveauxNon	Taille communauté: 181	
Rouendanslarue		
BFMTV	Taille communauté: 222	Taille communauté: 134
FranceInfo		
CNEWS		
76actu	Taille communauté: 204	Taille communauté: 200
paris_normandie		
DuPouvoirDachat	Taille communauté: 155	NA

4.4.2. Étude des relations entre les utilisateurs

Pour l'étude des relations entre les utilisateurs, nous avons étudié les phénomènes d'influence. Sous une forme plus algorithmique, il s'agit de découvrir les hubs et autorités dans le graphe des retweets en utilisant l'algorithme HITS (*Hyperlink-Induced Topic Search*) qui calcule deux scores pour chaque nœud, appelés score de hub et score d'autorité, uniquement en fonction des liens présents entre les nœuds (Kleinberg, 1999). Des résultats montrent que HITS fournit un calcul précis des nœuds d'autorité et des hubs (Devi *et al*., 2014). Nous avons appliqué HITS sur chacune des communautés détectées par la méthode Louvain (voir le tableau 4). Les utilisateurs qui sont des autorités sont des médias nationaux et locaux et des institutionnels. Les utilisateurs qui ressortent comme hub ont la particularité de beaucoup retweeter.

Tableau 4. Scores hub et autorité et communautés détectées par l'algorithme Louvain

Graphe de la relation retweet							
Autorités				**Hubs**			
Utilisateur	**Score**	**Rang**	**Communauté Louvain**	**Utilisateur**	**Score**	**Rang**	**Communauté Louvain**
BFMTV	1	1		YohannCrn	1	1	
franceinfo	0.682	2		bah_9	0.133	2	
Prefet76	1	1		Quinsolo	1	1	
gouvernementFR	0.298	2		lau68951920	0.661	2	
AiphanMarcel	1	1		MetropoleRouen	1	1	
76actu	1	1		DuPouvoirDachat	1	1	
raphtual	0.354	2		pythoncxde	0.350	2	

L'algorithme PageRank est aussi largement utilisé pour mesurer la centralité des utilisateurs. Nous avons ainsi étudié l'influence des utilisateurs dans leur communauté. En reprenant les communautés détectées par Louvain, nous avons calculé, communauté par communauté, le score PageRank de chaque utilisateur. La figure 3 reprend les communautés détectées avec le même code couleur que dans le tableau 3, la taille des nœuds est proportionnelle au score de Page Rank obtenu. Les nœuds ayant un score Page Rank élevé sont considérés comme autoritaires, d'ailleurs les comptes médiatiques (`BFMTV` et `FranceInfo`) sont les plus retweetés. D'un autre côté, `DuPouvoirDachat` est l'un des comptes qui retweetent le plus et il est vu comme un hub.

L'application de deux algorithmes permet aux chercheurs en Sciences Humaines et Sociales d'affiner leur interprétation dans les relations entre utilisateurs. Par exemple, l'utilisateur `76actu` qui est un média local est trouvé comme autoritaire avec l'algorithme HITS mais ne paraît pas comme un nœud central avec Page Rank dans sa communauté.

4.5. Analyses microscopiques : construction d'un discours de responsabilité institutionnelle dans une situation de crise

L'objectif de cette analyse est de voir comment les institutions s'expriment sur Twitter de manière responsable, comment elles incarnent la responsabilité d'information et de protection vis-à-vis des citoyens. Pour cela, le corpus Lubrizol a été restreint

Figure 3. Centralités intra-communautaires calculées après l'application de l'algorithme Louvain

pour étudier le discours des institutions (État, ministère, Préfet, académie, organismes publics). Pour se faire, l'équipe en Sciences Humaines et Sociales a sélectionné 23 comptes catégorisés par « institutions » et « organismes », le hashtag `#lubrizol` et le mot lubrizol réduisant alors le nombre total de tweets à 665.

Après une analyse qualitative en double aveugle, ces tweets ont été classés, suivant leur contenu, par modalité énonciative : « information », « conseil » et « obligation ». La figure 4 montre deux exemples de tweets catégorisés, avec les extraits de texte qui ont contribué à cette classification.

La figure 5 présente une synthèse de quelques utilisateurs par modalité énonciative en identifiant aussi les hubs et les autorités. À partir de ces résultats, les chercheurs en Sciences Humaines et Sociales ont fait l'analyse suivante : l'État, *via* ses différents acteurs institutionnels, semble chercher à construire une image « responsable » grâce aux informations et aux conseils apportés, faisant appel ainsi à la responsabilité du citoyen pour adopter une conduite appropriée. Le discours institutionnel porte assez peu sur l'obligation puisque le citoyen responsable n'a pas besoin de subir des injonctions : il agit en bon citoyen en connaissance de cause, en fonction des informations et des conseils reçus. Certains ont accusé l'État d'irresponsabilité en raison d'une stratégie de gestion de la peur, consistant à communiquer pour éviter la panique, tout en minimisant certains risques pour éviter l'affolement de la population.

Figure 4. Exemples de tweets classifiés

	% (nb de tweets contenant ce type de discours /nb de discours par compte)		
	Information	Conseil	Obligation
Académie de Rouen	60,42	35,42	4,17
ARS Normandie	72,92	18,75	8,33
Assemblée nationale	80,00	20,00	0,00
Atmo France	69,23	23,08	7,69
Gouvernement	85,37	9,76	4,88
Métropole Rouen Ndie	68,91	23,53	7,56
Ministère des Solidarités et de la Santé	82,14	16,07	1,79
Ministère de l'Intérieur	54,55	36,36	9,09
Préfet de la Seine-Maritime	70,75	25,47	3,77
Préfet de la Somme	93,33	6,67	0,00
Région Normandie	80,00	10,00	0,00

% (nb de tweets contenant ce type de discours /nb de discours par classe)		
Information	Conseil	Obligation
73,16	21,50	4,99

Hubs	Authorities

Figure 5. Pourcentage de tweets par modalité énonciative

La figure 6 met en perspective le nombre de tweets émis et catégorisés par modalité énonciative par rapport à des évènements significatifs qui se sont déroulés durant la période de l'étude. Nous notons que chaque évènement réel donne lieu à une émission de tweets.

Nous avons effectué à la fois une analyse quantitative et qualitative portant sur trois niveaux de granularité. L'étude macroscopique a mis en évidence l'interaction entre les utilisateurs qui communiquent sur Lubrizol. Elle a aussi mis en évidence le détournement du hashtag `#lubrizol` pour mettre en avant d'autres évènements. Et finalement, grâce à l'étude microscopique, nous avons étudié la communication institutionnelle lors d'une crise.

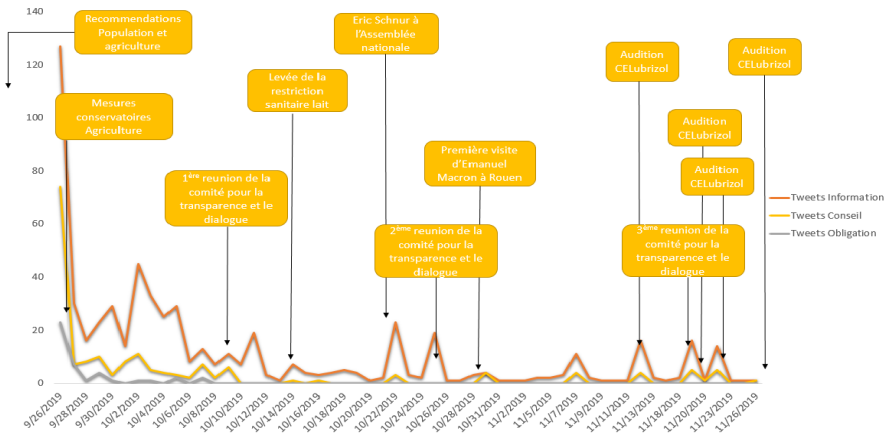


Figure 6. Série temporelle de l'évolution du nombre de tweets par modalité énonciative

5. Implémentation de la méthodologie sous une plateforme numérique

Le *notebook* Jupyter est un outil distribué sous licence BSD (*Berkeley Software Distribution*) utilisé pour l'analyse de données. Il permet aux *data scientists* de créer des scripts combinant du code, du texte et des interfaces graphiques (Perez, Granger, 2015). Des noyaux spécifiques à différents langages de programmation s'exécutent indépendamment et interagissent avec Jupyter, dont Python, R et Scala. Nous considérons le *notebook* Jupyter comme un environnement interactif qui nous permet de rassembler une description des données en entrée, de développer avec plusieurs langages de programmation dans un même noyau, et puis d'enregistrer et convertir les résultats vers d'autres formats que les fichiers structurés JSON tels que HTML et PDF.

Nous avons divisé l'implémentation en deux sections principales. La première consiste à la préparation pour effectuer les analyses, dans laquelle en utilisant le noyau Python, des tables sont créées sur une base PostgreSQL, après filtrage et regroupement des données brutes du corpus Lubrizol. La deuxième section comprend la partie analytique où les deux noyaux Python et R sont utilisés pour le calcul des indicateurs et des fréquences globaux, puis l'étude des graphes (communautés, centralité). À la fin nous utilisons une fonctionnalité Python pour visualiser les résultats sous forme de graphe à l'aide d'une connexion à Gephi. La figure 7 montre une capture du *notebook* Jupyter de l'analyse sur le corpus lubrizol, dans laquelle nous trouvons à gauche un panneau de navigation qui montre les sections qui composent ce *notebook*. À droite se trouvent les codes écrits en Python ou R correspondant aux sections du *notebook*.

D'autre part, une interface front-end est développée afin de fournir une visualisation des résultats aux *data analysts*/experts métiers, ce qui leur permet d'interpréter et de valider la méthodologie adaptée. Cette interface est développée en JavaEE, et

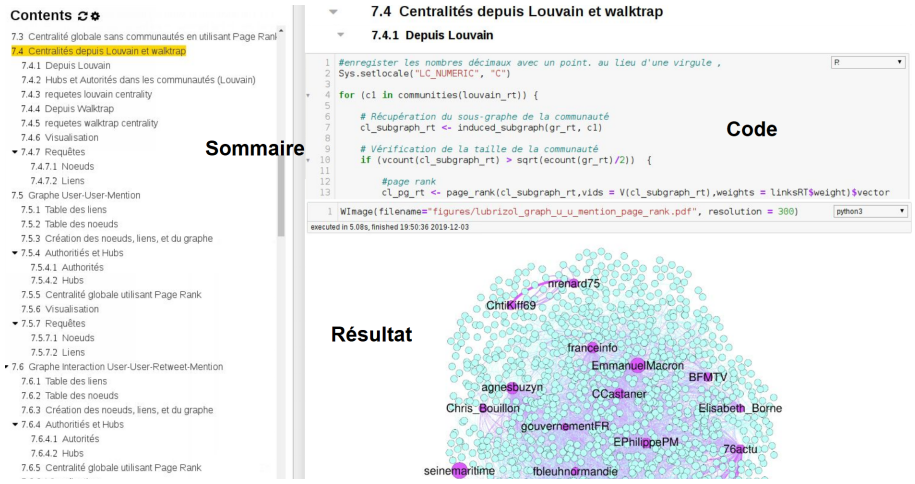


Figure 7. Le notebook Jupyter sur l'étude du corpus Lubrizol

les langages HTML, CSS et JQuery. Pour obtenir les résultats, deux méthodes sont possibles. La première consiste à extraire des séries temporelles d'un élément particulier de la base TimescaleDB (par exemple `#lubrizol`, `@Prefet76`, etc.). La deuxième méthode consiste à extraire le TOP k pour une catégorie d'éléments de la base (par exemple hashtag, mention, etc.). Ces deux méthodes utilisent les données enregistrées dans la base TimescaleDB qui est alimentée par la couche Speed décrite dans la section 3.

La figure 8 représente une série temporelle qui montre l'évolution du nombre des tweets par type (original, retweet, reply et quote), le *data scientist* peut aussi choisir de ne visualiser qu'un seul type, en fonction d'un intervalle de temps. Cet indicateur permet de suivre en temps réel l'émission de tweets sur une collecte et de voir si un évènement a lieu. Le même type d'indicateur est fourni pour suivre l'évolution en temps réel des hashtags.

6. Conclusion

Dans cet article, nous avons présenté notre démarche de travail et nos résultats d'analyse des tweets en l'appliquant à l'incendie de l'usine Lubrizol. Nos analyses ont été opérées à différents niveaux de granularité afin d'extraire des indicateurs significatifs. Nos analyses ont pour objectif de comprendre les interactions entre les utilisateurs, les communautés avec les utilisateurs influents, quels utilisateurs sont considérés comme des utilisateurs émettant une information fiable (autorité) et quels sont les utilisateurs qui relaient l'information (hub). Un focus sur le discours de responsabilité des institutions a été réalisé avec une approche qualitative, nous avons montré que la plupart des tweets étaient des tweets d'information.

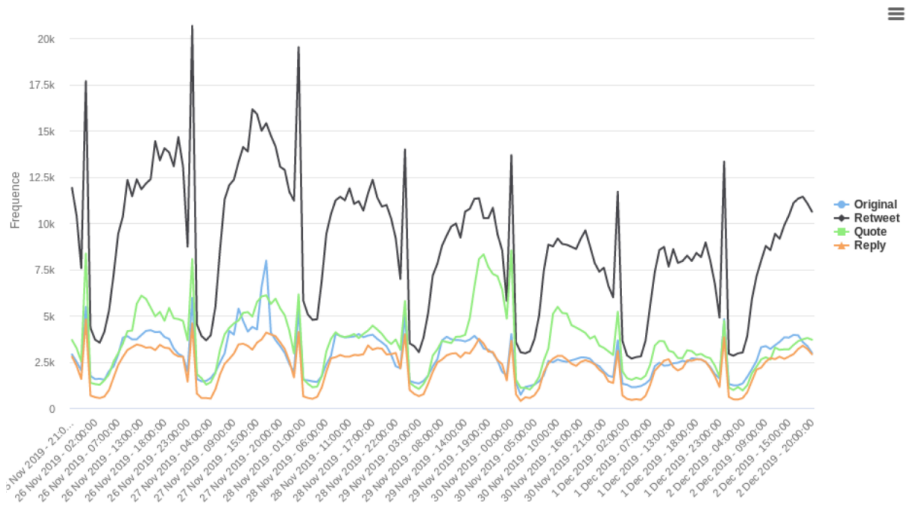


Figure 8. Interface front-end montrant l'évolution du nombre des tweets par intervalle d'une heure en fonction du temps

Dans la suite de notre travail, nous voulons détecter l'existence de signaux faibles. Un signal faible est une information d'alerte précoce, de faible intensité, pouvant être annonciatrice d'une tendance ou d'un évènement important (Ansoff, 1975). Le challenge porté par les signaux faibles est qu'ils sont difficiles à détecter et faciles à négliger, l'outillage algorithmique classique n'étant pas alors suffisant pour les détecter. C'est pourquoi nous voulons utiliser comme cadre théorique les « Graphlets » (Pržulj *et al.*, 2004). Les graphlets sont de petits (2 à 5 nœuds) sous-graphes induits connectés. Notre hypothèse est qu'ils sont une indication d'un début d'information précoce comme le début de la formation d'une communauté ou un mouvement agglomératif autour d'un hashtag.

Remerciement Ce travail est soutenu par le programme « Investissements d'Avenir », projet ISITE-BFC (contrat ANR-15-IDEX-0003). Le projet Cocktail est piloté scientifiquement par Gilles Brachotte, laboratoire CIMEOS EA-4177, Université de Bourgogne.

Bibliographie

Aiello L. M., Petkos G., Martin C., Corney D., Papadopoulos S., Skraba R. *et al.* (2013). Sensing trending topics in twitter. *IEEE Transactions on Multimedia*, vol. 15, nº 6, p. 1268–1282.

Ansoff H. I. (1975). Managing Strategic Surprise by Response to Weak Signals.

Azaza L., Leclercq É., Savonnet M. (2019). Modèle de réseaux multiplexes pour l'étude de l'influence sur twitter. In *Informatique des organisations et systèmes d'information et de décision (INFORSID)*, p. 255–270.

Basaille I., Kirgizov S., Leclercq É., Savonnet M., Cullot N. (2016). Towards a twitter observatory: A multi-paradigm framework for collecting, storing and analysing tweets. In *Conference on research challenges in information science (RCIS)*, p. 1–10.

Blondel V., Guillaume J.-L., Lambiotte R., Lefebvre E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, vol. 2008, n° 10, p. P10008.

Devi P., Gupta A., Dixit A. (2014). Comparative study of hits and pagerank link based ranking algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, vol. 3, n° 2, p. 5749–5754.

Fedoryszak M., Frederick B., Rajaram V., Zhong C. (2019). Real-time event detection on social data streams. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining*, p. 2774–2782.

Gillet A., Leclercq E., Cullot N. (2019). Lambda architecture pour une analyse à haute performance des données des réseaux sociaux. In *Informatique des organisations et systèmes d'information et de décision (INFORSID)*, p. 223–238.

Kleinberg J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, vol. 46, n° 5, p. 604–632.

MacEachren A. M., Robinson A. C., Jaiswal A., Pezanowski S., Savelyev A., Blanford J. *et al*. (2011). Geo-twitter analytics: Applications in crisis management. In *25th international cartographic conference*, p. 3–8.

Marz N., Warren J. (2015). *Big Data: Principles and best practices of scalable real-time data systems*. Manning.

Mora-Cantallops M., Sánchez-Alonso S., Visvizi A. (2019). The influence of external political events on social networks: The case of the brexit twitter network. *Journal of Ambient Intelligence and Humanized Computing*, p. 1–13.

Öztürk N., Ayvaz S. (2018). Sentiment analysis on twitter: A text mining approach to the syrian refugee crisis. *Telematics and Informatics*, vol. 35, n° 1, p. 136–147.

Perez F., Granger B. E. (2015). Project jupyter: Computational narratives as the engine of collaborative data science. *Retrieved September*, vol. 11, n° 207, p. 108.

Pons P., Latapy M. (2005). Computing communities in large networks using random walks. In *International symposium on computer and information sciences*, p. 284–293.

Pržulj N., Corneil D. G., Jurisica I. (2004). Modeling interactome: scale-free or geometric? *Bioinformatics*, vol. 20, n° 18, p. 3508–3515.

Vasiliu L., Freitas A., Caroli F., McDermott R., Zarrouk M., Hürlimann M. *et al*. (2016). In or out? Real-time monitoring of Brexit sentiment on Twitter. *SEMANTiCS*.

Gestion de données complexes

Lacs de Données : Tendances et Perspectives

Franck RAVAT[1], Yan ZHAO[1,2]

1. Institut de Recherche en Informatique de Toulouse, IRIT-CNRS (UMR 5505), Université Toulouse 1 Capitole, Toulouse, France
franck.ravat@irit.fr yan.zhao@irit.fr

2. Centre Hospitalier Universitaire (CHU) de Toulouse, Toulouse, France

RESUME. Le lac de données est actuellement présenté comme le composant essentiel pour l'analyse de mégadonnées. Dans cet article, nous résumons la définition, l'architecture fonctionnelle et les différents axes de recherche associés aux lacs de données de (Ravat, 2019) .

Mots-clés : Lac de données, Architecture, Métadonnées.

Keywords: data lake, architecture, metadata

À l'ère des mégadonnées (*Big Data*), l'analyse de données volumineuses, véloces et variées nécessitent des architectures adaptées pour l'intégration, le stockage et la restitution. Les entrepôts de données (ED) définis dans le cadre de la *Business Intelligence* (BI), ne sont plus adaptés pour les raisons suivantes : (i) seuls les besoins exprimés dès le début de la phase de conception peuvent être satisfaits ; (ii) toutes les données sources ne sont pas intégrées ; (iii) le coût d'implantation et de maintenance d'un ED peut croître de façon exponentielle pour assurer de bonnes performances d'interrogation et d'analyse. Pour relever le défi de l'analyse de mégadonnées, (Dixon, 2010) a été le premier à proposer le concept de Lac de Données (LD) dont l'objectif est de stocker toutes les données dans leur format natif. Cette définition imprécise n'est pas suffisante pour comprendre tous les enjeux de ce nouveau concept. Notre objectif est donc d'apporter une définition précise, de définir les composants d'une architecture générique de LD et d'aborder les futurs axes de recherche associés à ce nouveau concept.

Mêmes si les définitions ont évolué au fil du temps et à partir des retours d'expérience, ces dernières sont imprécises voire contradictoires. Pour être aussi complet que possible, nous définissons un LD au travers de ses entrées, ses processus de transformations, ses sorties et la gouvernance associée. Un LD est donc une solution d'analyse de mégadonnées qui (i) ingère et stocke des données brutes hétérogènes provenant de sources diverses dans leur format natif, (ii) permet de traiter et transformer ces données afin de répondre aux besoins d'analyse, (iii) fournit des accès aux données à un grand nombre d'utilisateurs pour des restitutions et/ou analyses

diverses (statistique, tableaux de bord interactif, analyse décisionnelle, exploration de données, apprentissage automatique etc.), et (iv) assure la qualité, la sécurité et le cycle de vie des données.

Afin de répondre aux lacunes et spécificités des différentes propositions, nous proposons une architecture fonctionnelle générique de LD basée quatre zones. Chaque zone contient des processus de traitement et un espace de stockage de données. La zone des données brutes permet d'ingérer en temps réel ou différé tous types de données dans leur format natif. La zone de traitements permet d'appliquer tous les processus de transformations et de calculs avec stockage de données intermédiaires afin de répondre à tout type d'analyse. La zone d'accès permet de préparer et stocker les données pour pouvoir appliquer une interrogation ou une analyse spécifique. Parallèlement, la zone de gouvernance, appliquée à toutes les autres zones, est chargée d'assurer la sécurité, la qualité, le cycle de vie, l'accès et la gestion des données à l'aide d'un système des métadonnées spécifique.

En complément de ces propositions, nous avons identifié différents axes de recherche prioritaires. Le premier axe de recherche est l'intégration d'un LD dans le système d'information d'une organisation pouvant déjà contenir un ED. La problématique de recherche est donc de savoir faire coexister deux systèmes ayant des objectifs distincts, comment faire circuler les données entre ces deux systèmes et quelles analyses complémentaires il est possible d'extraire. Le second axe de recherche est relatif à la définition et à la gestion des métadonnées (Ravat, 2019) afin de ne pas rendre les données d'un LD invisibles, incompréhensibles et inaccessibles. Ces métadonnées doivent informer de manière aussi complète que possible aussi bien sur les données que les processus de transformations (ingestion, traitement, diffusion) de ces données. Enfin, le troisième axe de recherche est relatif à la gouvernance des LD. Ces travaux de recherche doivent permettre de définir des politiques, des normes et des pratiques pour gérer les données de sources hétérogènes et les processus associés (transformation et analyse) afin d'assurer une utilisation efficace et sécurisée et une qualité fiable des résultats d'analyse. La gouvernance doit non seulement traiter des données mais également de tous les systèmes informatiques associés aux LD.

Bibliographies

Dixon, J. (2010) *Hadoop, and Data Lakes*, https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

Ravat, F., Zhao, Y. (2019) *Metadata management for data lakes,* East European Conference on Advances in Databases and Information Systems, Communications in Computer and Information Science. pp. 37–44. Springer International Publishing (2019)

Ravat, F., Zhao, Y. (2019) *Data lakes: Trends and perspectives*. Database and Expert Systems Applications - 30th International Conference, DEXA, Lecture Notes in Computer Science. pp. 304–313. Springer International Publishing (2019)

Modélisation de la dynamique des territoires : méta-données et lacs de données dédiés à l'information spatiale

Rodrique Kafando [1,3], **Rémy Decoupes** [1], **Lucile Sautot** [2], **Maguelonne Teisseire** [1]

1. *TETIS, Univ Montpellier, AgroParisTech, CIRAD, CNRS, INRAE, Montpellier, France*
   prenom.nom@inrae.fr
2. *AgroParisTech, Montpellier, France*
   lucile.sautot@agroparistech.fr
3. *Montpellier Méditerranée Métropole, France*

RÉSUMÉ. *La gestion efficace d'un lac de données nécessite un système de gestion de méta-données performant. De nombreux travaux se sont penchés sur cet aspect en proposant des solutions. Néanmoins, peu de travaux se sont intéressés aux lacs de données dédiés aux informations spatiales. Pourtant, cette dimension géographique est fondamentale dès lors que l'on souhaite explorer les différentes trajectoires de projets d'aménagement au sein d'un même territoire. Dans cet article, nous nous intéressons tout particulièrement à la mise en oeuvre d'un lac de données pour la métropole de Montpellier. La solution conceptuelle proposée s'adosse à la norme ISO 19115 pour décrire des méta-données spatiales qui est étendue dans le cadre des lacs de données. L'implémentation basée sur HDFS et GeoNetwork est présentée et discutée. Le code source est également mis à disposition de la communauté.*

ABSTRACT. *Data lake management requires an efficient metadata management system. Some works have already addressed this aspect in order to describe the datasets recorded and ensure their proper use. However, little work has been done on data lake dedicated to spatial information. However, geographical dimension is fundamental when we wish to explore the different trajectories of development projects within a territory. In this article, we are particularly interested in the implementation of a data lake for Montpellier metropolis. The proposed conceptual solution is based on the ISO 19115 standard to describe extended spatial metadata within the context of data lakes. The implementation based on HDFS and GeoNetwork is presented and discussed.*

MOTS-CLÉS : *Lac de données spatial, Données hétérogènes, Dynamique Territoriale*

KEYWORDS: *Spatial data lake, Heterogeneous data, Territorial Dynamic*

1. Introduction

Selon (Albino *et al*., 2015), les villes intelligentes sont définies comme des villes équipées en hautes technologies, qui connectent les habitants, les informations et les éléments urbains afin de créer une ville durable, un contexte économique compétitif et innovant, et une meilleure qualité de vie. Ces dernières années, nous avons constaté une croissance exponentielle des nouvelles technologies et des services associés en lien avec les villes intelligentes (Kitchin, 2014 ; Al Nuaimi *et al*., 2015). L'ensemble de ces services génèrent un grand volume de données, qui caractérisent, d'un point de vue global, l'évolution et le comportement du territoire.
Dans un tel contexte, les travaux présentés sont issus de la collaboration d'un laboratoire de recherche pluridisciplinaire avec Montpellier Méditerranée Métropole (3M). Le principal besoin exprimé par les utilisateurs est d'arriver à explorer sémantiquement de grandes quantités de données disponibles au sein de leur organisation. Parmi ces données, certaines sont produites par les citoyens, d'autres par les différents services de la métropole et des municipalités associées (transport, tourisme, etc.). Il est donc difficile d'avoir une vue d'ensemble sur les informations à disposition.

Le principal inconvénient des outils existants est la difficulté pour les utilisateurs d'explorer de manière flexible un ensemble de données massives et hétérogènes. Plus précisément, les entrepôts de données (Devlin, Cote, 1996) sont trop rigides pour permettre aux utilisateurs de construire de nouvelles analyses qui n'auraient pas été prévues (Madera, Laurent, 2016). Pour résoudre ce problème, les lacs de données (Dixon, 2010) représentent un nouveau mode de gestion des données, avec un stockage total ou partiel des éléments associés (données et méta-données). Pour ces nouveaux systèmes, dont la théorisation est récente, il y a eu peu de travaux méthodologiques sur la conception de ces infrastructures de données, considérant qu'elles requièrent essentiellement des compétences techniques. Confrontés à la mise en place d'un lac de données en conditions réelles dans le contexte d'une ville intelligente, nous sommes amenés à contredire cette hypothèse. Dans cet article, nous nous intéresserons ainsi à la conception et à l'implémentation d'un lac de données, en partant d'une masse de données hétérogènes avec une forte composante spatiale provenant de notre cas d'étude. En nous inspirant de travaux précédents sur la normalisation de données spatiales (ISO/TC 211, 2019) et sur les lacs de données (Ravat, Zhao, 2019 ; Madera, Laurent, 2016 ; Sawadogo *et al*., 2019), nous développons une méthodologie de conception dédiée. Le code développé de notre proposition est mis à disposition de la communauté. Nous démontrons que les infrastructures type lac de données ne sont pas réservées aux experts mais peuvent être proposées à d'autres utilisateurs à la condition de leur fournir une interface adaptée.

L'article est organisé comme suit. Dans la Section 2, nous présentons les définitions et les travaux relatifs à la conception de lac de données et à la gestion de l'information spatiale. Les données utilisées dans le cadre de la collaboration avec 3M sont détaillées dans la Section 3. La méthodologie proposée est présentée dans la Section 4, suivie par une description de l'implémentation dans la Section 5. La Section 6 conclut ce travail avec la discussion des résultats et la présentation de travaux à venir sur une solution de lac de données spatiales.

2. Etat de l'art

Plusieurs systèmes de gestion et de stockage de données ont émergé pour supporter le Big Data (McAfee *et al*., 2012). Parmi eux, nous pouvons citer les bases de données NoSQL (Not Only SQL) (Bruchez, 2015), les entrepôts de données (Kimball, Ross, 2011 ; Phipps, Davis, 2002) et les lacs de données (Russom, 2017 ; Hai *et al*., 2016).

2.1. Entrepôts de données et lac de données

Les entrepôts de données ont été conçus comme une optimisation des bases de données relationnelles pour l'exécution de requêtes analytiques et sont utilisés comme support à la prise de décision dans les organisations. Les modèles conceptuels des entrepôts de données sont basés sur les concepts suivants : les faits (et mesures), les dimensions, les hiérarchies et les membres (Kimball, Ross, 2013). De fait, concevoir un entrepôt de données revient à définir l'espace des tableaux croisés possibles, qui vont être construits par les utilisateurs pour explorer les données. Les entrepôts de données permettent une exploration facilitée de volumineux jeux de données par les utilisateurs. Mais la mise en place d'un entrepôt de données implique une normalisation des données entrantes issues de sources variées (cette normalisation pouvant être automatisée via un ETL). Malgré quelques propositions intéressantes (voir par exemple (Oukid *et al*., 2016) et (Minati *et al*., 2006)), l'intégration de documents et d'images satellitaires dans le même entrepôt de données reste une tâche complexe. Ainsi, la mise en place d'un entrepôt de données nécessite un long processus de préparation des données.

La définition d'un lac de données a été proposée par (Dixon, 2010). Une comparaison détaillée avec les entrepôts de données a été réalisée dans (Madera, Laurent, 2016) puis reprise dans (Sawadogo *et al*., 2019). Les lacs de données sont une solution récente qui a été développée pour répondre à la gestion des Big Data pour lesquelles les entrepôts de données montraient quelques faiblesses. Le principal problème rencontré avec les entrepôts de données est la gestion de données de natures hétérogènes. Un lac de données est une structure de stockage de données massives, qui intègre les données en provenance de différentes sources dans leur format natif, sans qu'il soit nécessaire de réaliser un traitement (Russom, 2017 ; Hai *et al*., 2016). Selon (Sawadogo *et al*., 2019). Un lac de données est un système évolutif de stockage et d'analyse de données, stockées dans leur format natif, destiné

à des spécialistes tels que des statisticiens, des analystes et des "data scientists". Les principales composantes et caractéristiques des lacs de données sont :

– un catalogue de méta-données qui facilite l'accès aux données et en assure la qualité,

– des outils de gestion des données,

– l'accessibilité aux utilisateurs,

– l'évolution possible des données,

– l'ingestion des données de toute nature,

– une organisation logique et physique.

Etant une nouvelle technologie Big Data, les lacs de données ont été étudiés dans de nombreux articles. En raison de leur capacité à gérer de larges volumes de données, structurées et non structurées, une étude exploratoire a été réalisée pour mieux comprendre l'utilisation des lacs de données dans le contexte industriel (Llave, 2018). Dans (Giudice *et al.*, 2019 ; Mehmood *et al.*, 2019), de nouvelles architectures de lac de données ont été conçues afin d'extraire des informations pertinentes d'un ensemble de données hétérogènes, en se basant sur les sources de ces données. Dans (Quix *et al.*, 2016), un système de gestion de méta-données générique et extensible pour les lacs de données (Generic and Extensible Metadata Management System for Data Lakes, GEMMS) a été développé, en premier lieu pour extraire des méta-données des sources et en second lieu, pour enrichir les sources de données en utilisant des informations sémantiques venant à la fois des données et des méta-données. De nombreux système de gestion de méta-données ont été ainsi proposés par la communauté, mais il reste encore des défis à relever dont en particulier la mise en lien sémantique des données (Nargesian *et al.*, 2019).

En nous basant sur ces travaux, nous définissons un lac de données comme une structure de stockage composée de jeux de données, ayant les caractéristiques précédemment citées et celles décrites dans la Section 4.

2.2. Information géographique

Plusieurs définitions ont été proposées pour le concept de territoire selon le domaine étudié. Dans (Moine, 2006), le territoire est considéré comme étant un système complexe et évolutif qui associe un ensemble d'acteurs, d'une part, et d'autre part, l'espace géographique que ces acteurs utilisent, développent et gouvernent. Dans (Simone *et al.*, 2018), les auteurs quant à eux, considèrent que le territoire est un ensemble composé de trois dimensions : l'espace géographique, le temps et les relations sociales. Ils définissent le territoire comme étant un système complexe situé dans un espace géographique spécifique émergeant de la co-évolution d'un ensemble de processus hétérogènes (anthropologico-culturel, relationnel, cognitif et économico-productif) qui caractérise cet espace d'une manière unique et non répétitive.

Tout en prenant en compte les définitions proposées dans l'état de l'art, nous considérons que le territoire est :

– un ensemble d'acteurs physiques et/ou juridiques. Physique dans le sens où il est habité par un ou plusieurs groupes de personnes interagissant les uns avec les autres, et juridique au sens où il est composé de plusieurs organisations politiques, économiques, etc.

– décrit par un ensemble d'informations géographiques, à savoir des entités spatiales, thématiques et temporelles qui interagissent entre elles. Ces informations évoluant dans le temps et dans l'espace.

Dans cette étude, nous nous focalisons principalement sur les informations géographiques produites et gérées au niveau de la Métropole de Montpellier qui est notre zone d'étude. Notre proposition est basée sur la norme ISO 19115 (ISO/TC 211, 2019) dédiée aux données spatiales (identification, étendue, qualité, contenu, référence géographique, etc.).

3. Données et utilisateurs du lac de données

Les jeux de données utilisés dans le cadre de notre étude sont constitués entre autres d'images satellites, de documents textuelles, de couches vectorielles et autres données telles que les données de transports, d'urbanisation, d'agriculture, de commerce, etc. Elles proviennent de sources différentes :

– la plate-forme opendata de la Métropole de Montpellier [1]. Elle regroupe un ensemble de données produites par la Métropole de Montpellier et qui sont mises à la disposition du grand public. La liste exhaustive des liens se retrouve dans le fichier datasources.csv présent dans le dépôt logiciel de notre implémentation (https://github.com/aidmoit/collect/blob/master/input/datasources.csv), accédé le 2020-02-19. Ces jeux de données sont publiées sous licence "Open Data Commons Open Database License" (ODbL).

– le web : nous avons constitué des corpus de données textuelles à partir du web. Les corpus sont construits en tenant compte des thématiques abordées que nous souhaitons étudier dans la phase de mise en relation.

– OpenStreetMap : nous a permis d'obtenir les étendues spatiales des lieux de notre cas d'étude (les communes de la métropole de Montpellier), accédé le 2020-02-19.

Notre solution peut être exploitée par deux types d'utilisateurs. Tout d'abord, les utilisateurs 'grand public' ou tout utilisateur, le système leur permet d'explorer et de récupérer des données présentes dans le lac à travers l'interface web offert par GeoNetwork sans avoir besoin de compétences sur l'exploitation des lac de données. Ensuite les utilisateurs expérimentés, qui en plus de l'exploration peuvent effectuer

1. Open Data 3M : http://data.montpellier3m.fr/

des traitements et des analyses directement sur le lac de données en utilisant des outils comme Apache Spark.

4. Architecture du lac de données spatiales

Dans cette section, nous proposons une vue générale de l'architecture d'un lac de données spatiales, avec pour objectif de fournir un guide pour reproduire une telle architecture. L'architecture proposée est prévue pour stocker les données produites et utilisées par la métropole de Montpellier. Ce cas d'étude implique plusieurs contraintes :

– la dimension spatiale des jeux de données et les analyses spatiales réalisées sont un élément important. Nous avons notamment besoin de stocker des images satellites.

– le système proposé doit être inter-opérable avec d'autres systèmes d'information, au niveau local, national et européen.

– les utilisateurs souhaitent explorer le lac de données pour y trouver des données pertinentes et découvrir de nouvelles connaissances.

L'architecture proposée est composée de trois parties principales : la section data, la section metadata et la section intermetadata. La section data, le coeur de la structure de stockage, est basée sur Hadoop Distributed File System (HDFS) (Shvachko *et al.*, 2010). Le choix de HDFS est motivé d'une part, par le fait qu'il permet de stocker les données dans leur format natif (contrairement au système de stockage clé-valeur), et d'autre part, de sa distributivité. Avec HDFS, il est possible détendre (parallèlement) plus facilement la capacité de stockage en cas de besoin et aussi d'effectuer des calculs distribués.

La section metadata est un catalogue de données (Lamb, Larson, 2016), qui décrit les données stockées dans le lac de données. La section intermetadata est une partie de la section metadata. Elle permet le stockage de relations entre jeux de données riches sémantiquement.

HDFS est un système performant pour le stockage de données massives et hétérogènes, mais ne peut pas être utilisé tel quel par nos utilisateurs. Les utilisateurs du lac de données ont besoin d'explorer le lac de données afin de trouver les jeux de données les plus pertinents vus leur requête, et éventuellement de découvrir de nouveaux jeux de données. Ces fonctionnalités (exploration, requêtage, découverte) sont supportées par le catalogue de données, qui offre une interface graphique simple pour accéder aux méta-données descriptives du contenu du lac de données.

Le modèle conceptuel que nous proposons est une extension de la norme ISO 19115 (ISO/TC 211, 2019). Cette norme inclut une description spatiale des données et sert de base à plusieurs profiles de méta-données (INSPIRE, Dublin Core) utilisés habituellement par les institutions publiques.

La FIGURE 1 est une vue générale du modèle conceptuel étendu proposé. Dans cette figure, la classe représentée en blanc est directement issues de la norme ISO

19115 (FIGURE 1) et les classes représentées en jaune constituent nos ajouts. Afin que les modèles restent lisibles, nous avons fait le choix de ne représenter que la classe principale de chaque package.

Dans la section data (FIGURE 2), nous définissons un lac de données comme un ensemble de ressources. Une ressource peut être un service (voir la norme ISO 19115) ou une série de données. Une série de données est composée d'un ou plusieurs jeux de données, qui partagent une caractéristique. Un jeu de données est une collection de données identifiables. Trois types de jeux de données ont été définis : document, vecteur et raster.

La section metadata décrit les fiches de méta-données associées à chaque ressource (voir FIGURE 3). Une fiche de méta-données est composée de :

– une identification (obligatoire) qui permet la différenciation des ressources par l'utilisateur ;

– une représentation spatiale (optionnelle), un système de coordonnées de référence (optionnel) et une emprise spatiale et/ou temporelle (optionnelle). Ces trois éléments permettent de décrire la spatialité de la ressource ;

– une description du contenu de la ressource (optionnelle) ;

– une généalogie (optionnelle), qui explique comment la ressource a été obtenue ;

– un ou plusieurs liens vers des ressources associées ;

– un système de référence (optionnel), qui identifie les systèmes de références spatiaux, temporels et paramétriques utilisés par cette ressource ;

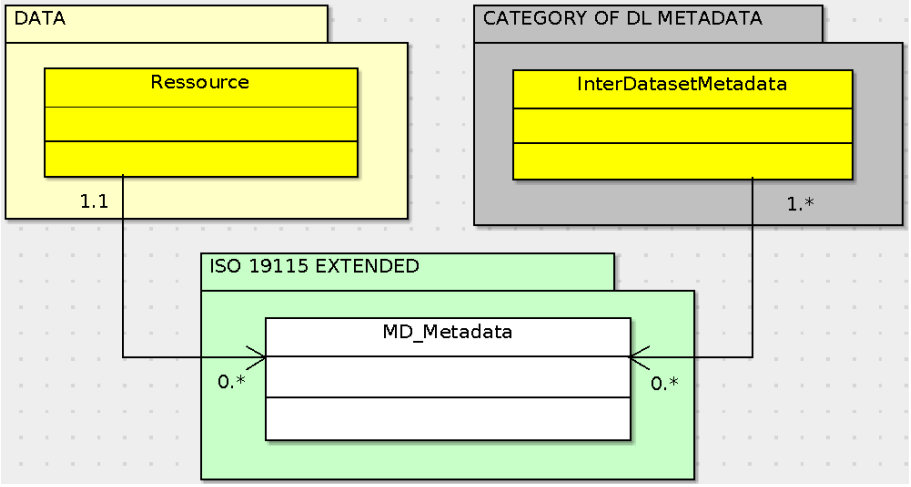
– une emprise, qui décrit l'emprise temporelle et spatiale de la ressource.

Enfin, la section intermetadata (voir FIGURE 4) décrit les relations entre les jeux de données et permet à l'utilisateur d'avoir une visibilité sur les données liées à sa requête initiale. Quatre types de relations ont été proposés, basés sur (Sawadogo *et al.*, 2019) : parenté, inclusion, similarité et regroupement thématique. Le modèle conceptuel proposé dans son ensemble permet de prendre en compte non seulement l'intégration des méta-données de données spatiales, mais aussi tout type de données stockées dans le lac.

5. Implémentation pour la Métropole de Montpellier : 3M (Montpellier Méditerranée Métropole)

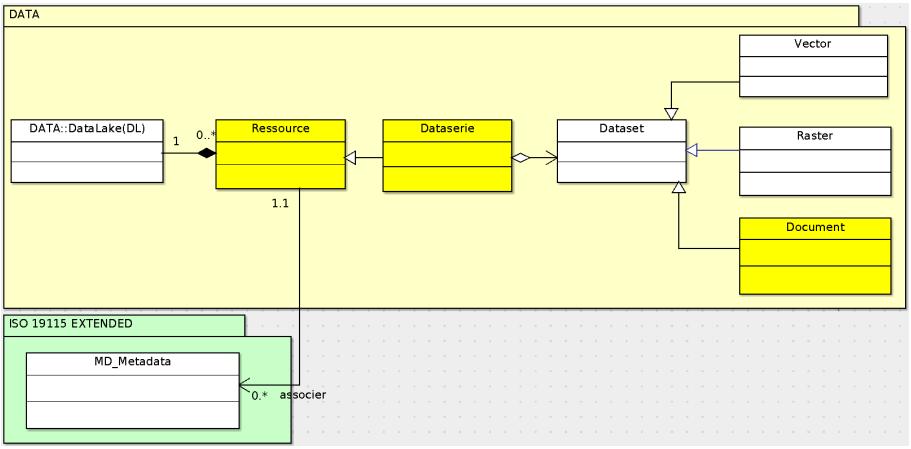
L'objectif de cette section est de présenter la mise en œuvre du lac de données pour la Métropole de Montpellier [2].

2. Montpellier Méditerranée Métropole (3M) : https://www.montpellier3m.fr/



Les classes blanches viennent de (ISO/TC 211, 2019), les classes jaunes ont été ajoutées à la norme.

Figure 1. Vue générale du modèle conceptuel proposé

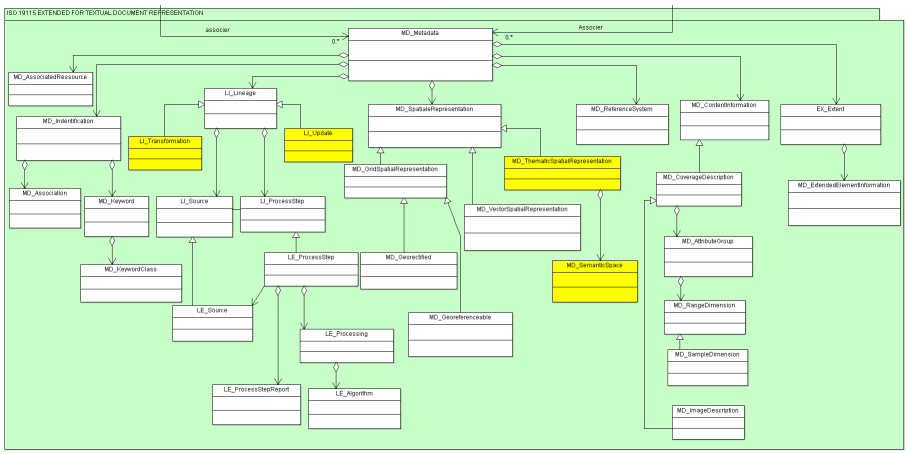


Les classes blanches viennent de (ISO/TC 211, 2019), les classes jaunes ont été ajoutées à la norme.

Figure 2. Section Data

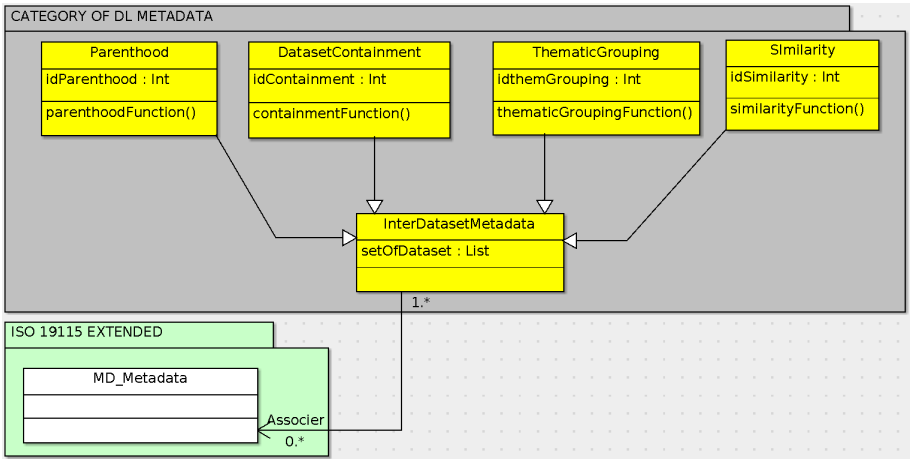
5.1. Présentation de l'infrastructure système

Comme présenté dans la précédente section, un lac de données est composé de deux sous-systèmes. La partie donnée est assurée par la mise en place d'un système de fichiers distribués. Cette partie est enrichie par un gestionnaire de méta-données qui constitue le deuxième composant du lac de données.
Dans notre implémentation, la partie données repose sur le système de fichiers distribués HDFS (Hadoop Distributed File System), utilisant la technologie du projet



Les classes blanches viennent de (ISO/TC 211, 2019), les classes jaunes ont été ajoutées à la norme.

Figure 3. Section metadata



Les classes blanches viennent de (ISO/TC 211, 2019), les classes jaunes ont été ajoutées à la norme.

Figure 4. Section intermetadata

HADOOP [3]. Notre ensemble de serveurs HDFS, appelé cluster HDFS, est constitué de trois machines ayant des rôles différents : deux nœuds de stockage (appelé datanode dans la terminologie Hadoop) et un nœud maître (appelé namenode). Ce dernier gère les interactions avec les utilisateurs ainsi que la distribution et la réplication des données sur les datanodes.

Malgré les différentes fonctionnalités offertes à l'utilisateur, le namenode HDFS ne propose ni de système d'indexation de données ni de moteur de recherche. Pour

3. Hadoop : http://hadoop.apache.org/

répondre à ce besoin, l'administrateur d'un lac de données doit avoir recourt à un outil de gestion de méta-données de type ElasticSearch, construit sur le projet Apache Lucene (Chen *et al.*, 2017) et (John, Misra, 2017). Nous proposons d'utiliser l'outil GeoNetwork [4]. Cet outil open-source embarque un moteur de recherche Apache Lucene et a l'avantage d'implémenter le modèle de la norme ISO 19115. Ainsi, le serveur GeoNetwork sauvegarde les méta-données obligatoires et optionnelles, telles que décrites dans la précédente section, et conserve les liens permettant de télécharger les données stockées dans le cluster Hadoop via le namenode. Le moteur de recherche de GeoNetwork permet à l'utilisateur de faire des recherches croisées sur les trois dimensions : spatiales, temporelles et thématique. Le résultat de la recherche est une collection de jeux de données répondant à l'intersection des critères de la requête.

5.1.1. Insertion et indexation des données dans le lac de données

Comme illustrée par la FIGURE 5, l'insertion de jeux de données dans le lac de données se déroule en cinq étapes. Les deux premières étapes sont réalisées manuellement, les trois dernières étapes sont, elles, automatisées.
En effet, l'administrateur doit remplir un tableur au format CSV [étape 1]. Ses colonnes sont les méta-données telles que décrites dans la section précédente ainsi qu'un lien HDFS pour indiquer l'emplacement du jeu de données. Chaque ligne représente un jeu de données pour lequel l'administrateur doit compléter les méta-données.
Puis l'administrateur lance un programme, voir section "Accès aux logiciels et données de l'implémentation", depuis sa machine (ou un serveur du lac de données) [étape 2]. Au début de son exécution, le programme va lire [étape 3] et extraire les informations du fichier CSV. Puis le programme, télécharge les jeux de données et les insère dans le cluster HDFS [étape 4]. Enfin, le programme crée une fiche de méta-données de type ISO 19115 et l'insère dans GeoNetwork afin de bénéficier de son indexation et de son moteur de recherche [étape 5].

5.1.2. Découverte et accès aux jeux de données

L'utilisateur peut parcourir, découvrir, faire une requête et accéder aux jeux de données en utilisant le moteur de recherche de GeoNetwork. Les recherches peuvent être une combinaison de critères sur les trois dimensions :

– sémantique : basé sur les mots clés ou bien sur une recherche en texte plein sur le titre, le résumé ou la généalogie de la fiche de méta-données.

– spatialisée : en dessinant une emprise spatiale directement sur la carte afin de filtrer les jeux de données qui intersectent l'étendue géographique voulue.

– temporelle : filtrer sur les années, mois et jour.

4. GeoNetwork : application web de catalogue de données spatialisées. https://geonetwork-opensource.org/

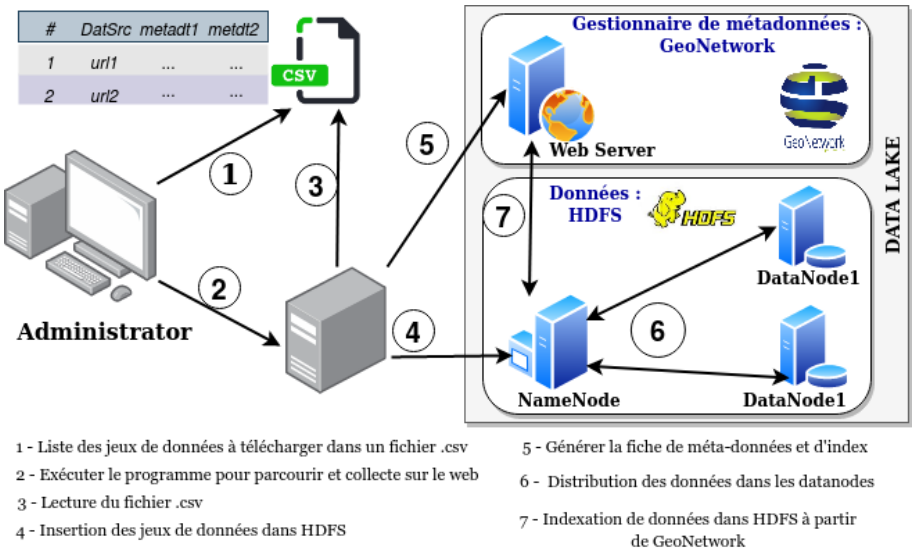


Figure 5. Insertion et indexation de jeux de données dans le lac de données

GeoNetwork retourne une collection de fiches de méta-données décrivant les jeux de données qui respectent les critères de recherche. En parcourant les jeux de données, l'utilisateur peut accéder à tous les fichiers de données stockés dans le cluster HDFS sans avoir besoin de connaître la syntaxe d'interrogation d'Hadoop (FIGURE 6).

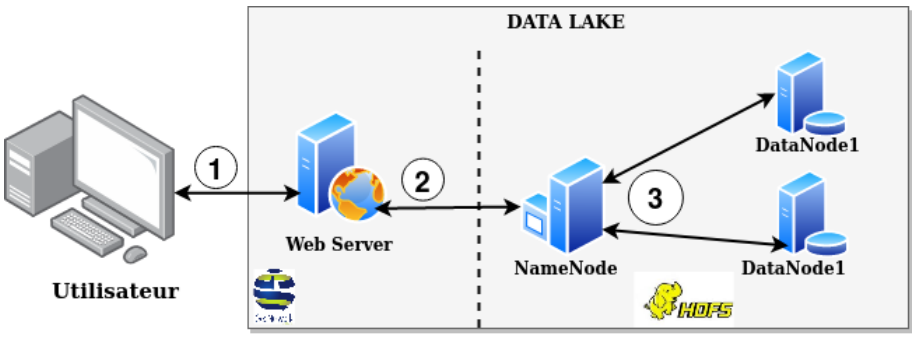


Figure 6. Recherche de jeux de données

5.2. Automatisation du déploiement du lac de données et de l'ajout de contenu

L'installation d'un cluster Hadoop peut s'avérer complexe. À des fins de reproductibilité, nous avons automatisé son installation, via les outils Vagrant [5] et Ansible [6].
Cette automatisation construit et configure quatre machines virtuelles dont les trois premières appartiennent au cluster Hadoop et la dernière héberge le serveur GeoNetwork. Ceci permet d'instancier facilement notre lac de données.
Nous avons aussi automatiser l'ajout de jeux de données ainsi que leur indexation. Cette automatisation prend, cette fois-ci, la forme d'un script python qui analyse les informations provenant d'un fichier CSV, y extrait des liens de téléchargement qui lui permettent de télécharger les jeux de données voulus. Ensuite, le script insère les fichiers de données dans le cluster HDFS sans les organiser dans une arborescence. Enfin, il crée des fiches de méta-données, les insère dans GeoNetwork afin de bénéficier de son moteur de recherche.

5.2.1. Déploiement automatique du cluster HDFS et du serveur GeoNetwork

L'ensemble du lac de données, c'est-à-dire le cluster HDFS et le serveur GeoNetwork, est déployé et maintenu grâce à l'utilisation de projets opensource notamment les suivants :

– Debian: Système d'exploitation utilisé par les 4 machines virtuelles. Nous avons utilisé la version 9 et non pas la dernière version à cause de problème de compatibilité avec la version java nécessaire à Hadoop et à GeoNetwork. La dernière version (version 10) de debian ne maintient plus cette version de java (version 9),

– VirtualBox comme hyperviseur,

– Vagrant comme système de gestion de configuration des machines virtuelles (système d'exploitation utilisé, configuration réseau, script d'installation, ...),

– Ansible comme un outil de déploiement d'application et de gestion de configuration.

Les codes sources de ce projet peuvent être retrouvés dans la section 5.4. Grâce à ces dépôts logiciels, le cluster HDFS peut être déployé et configuré en quatre commandes et le serveur GeoNetwork en une commande.

De plus amples informations ou instructions techniques peuvent être retrouvées dans le fichier README.md du dépôt logiciel de notre projet. Si les variables par défaut, proposé par le dépôt logiciel, sont conservées, le cluster HDFS peut être accessible de manière graphique en se connectant à son serveur web à l'adresse http://namenode:9870 (ou http://10.0.0.10:9870). D'autres informations, telles que la santé du cluster, ou l'accès aux logs ou bien encore l'accès au système de fichiers HDFS peuvent être aussi retrouvé via cette interface. Le serveur GeoNetwork est

5. Vagrant : https://www.vagrantup.com/
6. Ansible : https://www.ansible.com/

quant à lui accessible à l'adresse http://aidmoit-geonetwork:8080/geonetwork (http://10.0.0.9:8080/geonetwork).

5.2.2. Ajout de données dans le lac de données

L'ajout de données dans le lac de données a, lui aussi, été automatisé. Deux scripts, en python et en R, ont été écrits. La complexité qu'induit le développement d'un outil basé sur deux langages de programmation différents a été motivé par les couvertures de fonctionnalités des librairies de chaque langage. En effet, python offre des libraires remarquables pour interagir avec HDFS alors que R propose des modules intéressants pour gérer des fiches de méta-données compatibles ISO 19115. Afin de faciliter l'utilisation de ces deux scripts, le programme R a été encapsulé dans le code python, permettant, ainsi, à l'administrateur, de ne lancer qu'un seul programme.

Comme mentionné auparavant, l'ensemble des fichiers sources est disponible dans la section 3. L'environnement requis pour faire fonctionner ces scripts a été décrit dans le fichier "requirement.txt" présent à la racine du dépôt du logiciel. Les instructions d'installation et de lancement sont, quant à eux, présentées dans le fichier README.md.

Le script principal écrit en python opère en cinq étapes. Premièrement, il extrait les informations contenu dans le fichier datasources.csv comme le fournisseur de données, le nom du jeu de données et les mots clés associés. Le script, dans une deuxième étape, parcourt le site web du fournisseur de données afin de créer un fichier json contenant l'ensemble des liens de téléchargement des jeux de données. Ensuite, tous les fichiers constituant les jeux de données sont téléchargés, puis enregistré dans le cluster HDFS, ce qui constitue les troisième et quatrième étapes. Enfin, le script R est lancé afin de créer des fiches de méta-données au standard ISO 19139 qui sont, ensuite, ingérées par GeoNetwork.

Les fichiers de données sont facilement récupérables à partir de GeoNetwork. En effet, le namenode du cluster HDFS offre une interface de programmation de type API REST (Application Programming Interface - REpresentational state transfer) permettant une abstraction complète des commandes HDFS.

5.3. Illustration d'une recherche d'un utilisateur

L'utilisateur de notre lac de données peut créer des requêtes complexes mélangeant les trois dimensions : spatiale, temporelle et sémantique. Ces requêtes se construisent à travers l'utilisation du moteur de recherche de GeoNetwork qui offre plusieurs méthodes de composition de recherche. En effet, la requête peut être élaborée via une combinaison de recherche en texte libre (sur les trois dimensions) et/ou par mot clés (aussi sur les trois dimensions) et/ou par le dessin d'une étendue spatiale sur une carte (uniquement dimension spatiale).

5.4. Accès aux logiciels et données de l'implémentation

5.4.1. Infrastructure système

L'infrastructure du lac de données, c'est-à-dire, le cluster HDFS et le serveur GeoNetwork, est instanciable à travers l'utilisation de quatre machines virtuelles. L'installation et le lancement de ces machines ont été automatisés. Le dépôt logiciel est le suivant : https://github.com/aidmoit/ansible-deployment. Les instructions d'utilisation sont décrites dans le fichier README.md du dépôt. Le numéro de commit utilisé pour notre implémentation est le suivant :
65de950a336ee2828cdb19db976b7946649c439c
Le dépôt est publié sous la licence GPL-3.

5.4.2. Logiciel et flux de traitement

L'ensemble des logiciels pour le téléchargement des données, leurs ajouts dans le cluster HDFS et leurs descriptions dans le GeoNetwork sont orchestrés par un script python. Toutes les ressources nécessaires à son exécution sont disponibles à travers ce dépôt : https://github.com/aidmoit/collect. Le numéro de commit utilisé pour notre implémentation est le suivant : da9f63f9287a191d7e8fd24884a731bae02e1034.

Les codes sont distribués sous la licence GPL-3. Ils exploitent deux paquets R : geometa (Blondel, 2019) et geonapi (Blondel, 2018) diffusés sous licence MIT. Enfin, les scripts utilisent les données d'OpenStreetMap [7], ces données sont publiées sous licence "Open Data Commons Open Database License" (ODbL).

6. Conclusion et Perspectives

Dans cet article, nous avons présenté une nouvelle méthodologie de conception et d'implémentation de lac de données spatiales. La principale contribution est l'introduction de la dimension spatiale dans le processus de conception de lac de données, basée sur un système de méta-données géographiques. Nous avons également montré que les lacs de données peuvent être orientés vers les utilisateurs finaux, ce qui est possible en mettant en place une interface de requêtes. Nos travaux futurs seront dédiés à l'analyse et à la mise en lien des données stockées dans le lac de données. Les questions à traiter seront :

1. Comment des données hétérogènes peuvent être liées sémantiquement pour une analyse des phénomènes spatio-temporels complexes qui ont lieu sur un territoire ?

2. Quelles méthodes originales de fouille de données faut-il utiliser pour analyser des données hétérogènes massives ?

Atteindre ces objectifs nous permettra d'une part de décrire des relations riches entre les données selon les thématiques et le contexte spatio-temporel, et d'autre part de contribuer à la description de l'évolution d'un territoire. En d'autres termes, le

7. OpenStreeMap: https://www.openstreetmap.org

concept de lacs de données spatiales devient un élément central dans le dispositif des villes intelligentes.

Remerciements

Ces travaux ont été partiellement financés par Montpellier Méditerranée Métropole et le Projet Songe (FEDER et Région Occitanie).

Bibliographie

Albino V., Berardi U., Dangelico R. M. (2015, janvier). Smart Cities: Definitions, Dimensions, Performance, and Initiatives. *Journal of Urban Technology*, vol. 22, n° 1, p. 3–21. Consulté sur http://www.tandfonline.com/doi/full/10.1080/10630732.2014.942092

Al Nuaimi E., Al Neyadi H., Mohamed N., Al-Jaroodi J. (2015). Applications of big data to smart cities. *Journal of Internet Services and Applications*, vol. 6, n° 1, p. 25.

Blondel E. (2018, août). *geonapi: R interface to geonetwork api.* Zenodo. Consulté sur https://doi.org/10.5281/zenodo.1345102

Blondel E. (2019, octobre). *geometa: Tools for Reading and WritingISO/OGC Geographic Metadata in R.* Zenodo. Consulté sur https://doi.org/10.5281/zenodo.3524348

Bruchez R. (2015). *Les bases de données NoSQL et le BigData: Comprendre et mettre en oeuvre*. Editions Eyrolles.

Chen D., Chen Y., Brownlow B. N., Kanjamala P. P., Arredondo C. A. G., Radspinner B. L. *et al.* (2017, April). Real-time or near real-time persisting daily healthcare data into hdfs and elasticsearch index inside a big data platform. *IEEE Transactions on Industrial Informatics*, vol. 13, n° 2, p. 595-606.

Devlin B., Cote L. D. (1996). *Data warehouse: from architecture to implementation*. Addison-Wesley Longman Publishing Co., Inc.

Dixon J. (2010, octobre). *Pentaho, Hadoop, and Data Lakes.* Consulté sur https://jamesdixon.wordpress.com/2010/10/14/pentaho-hadoop-and-data-lakes/

Giudice P. L., Musarella L., Sofo G., Ursino D. (2019). An approach to extracting complex knowledge patterns among concepts belonging to structured, semi-structured and unstructured sources in a data lake. *Information Sciences*, vol. 478, p. 606–626.

Hai R., Geisler S., Quix C. (2016). Constance: An intelligent data lake system. In *Proceedings of the 2016 international conference on management of data*, p. 2097–2100.

ISO/TC 211. (2019). *Norme iso 19115-1:2014. geographic information - metadata - part 1: Fundamentals. technical report, international organization for standardization, 2019.* International Organization for Standardization.

John T., Misra P. (2017). *Data lake for enterprises: Lambda architecture for building enterprise data systems*. Packt Publishing.

Kimball R., Ross M. (2011). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. John Wiley & Sons.

Kimball R., Ross M. (2013). *The data warehouse toolkit: The definitive guide to dimensional modeling*. John Wiley & Sons.

Kitchin R. (2014). The real-time city? big data and smart urbanism. *GeoJournal*, vol. 79, nº 1, p. 1–14.

Lamb I., Larson C. (2016). Shining a light on scientific data: Building a data catalog to foster data sharing and reuse. *Code4Lib Journal*, nº 32.

Llave M. R. (2018). Data lakes in business intelligence: reporting from the trenches. *Procedia computer science*, vol. 138, p. 516–524.

Madera C., Laurent A. (2016). The next information architecture evolution: the data lake wave. In *Proceedings of the 8th International Conference on Management of Digital EcoSystems*, p. 174–180. ACM.

McAfee A., Brynjolfsson E., Davenport T. H., Patil D., Barton D. (2012). Big data: the management revolution. *Harvard business review*, vol. 90, nº 10, p. 60–68.

Mehmood H., Gilman E., Cortes M., Kostakos P., Byrne A., Valta K. *et al.* (2019). Implementing big data lake for heterogeneous data sources. In *2019 ieee 35th international conference on data engineering workshops (icdew)*, p. 37–44.

Minati L., Ghielmetti F., Ciobanu V., D'Incerti L., Maccagnano C., Bizzi A. *et al.* (2006). Bio-image warehouse system: Concept and implementation of a diagnosis-based data warehouse for advanced imaging modalities in neuroradiology. *Journal of Digital Imaging*, vol. 20.

Moine A. (2006). The territory as a complex system: an operational concept for land planning and geography (le territoire comme un systeme complexe: un concept opératoire pour l'aménagement et la géographie). *Esp. Géogr*, vol. 2, nº 35, p. 115.

Nargesian F., Zhu E., Pu K. Q., Miller R. J., Arocena P. C. (2019). Data lake management: Challenges and opportunities. , vol. 12, nº 12, p. 4.

Oukid L., Boussaid O., Benblidia N., Bentayeb F. (2016). Tlabel: A new olap aggregation operator in text cubes. *International Journal of Data Warehousing and Mining*, vol. 12, nº 4, p. 54-74.

Phipps C., Davis K. C. (2002). Automating data warehouse conceptual schema design and evaluation. In *Dmdw*, vol. 2, p. 23–32.

Quix C., Hai R., Vatov I. (2016). Metadata extraction and management in data lakes with gemms. *Complex Systems Informatics and Modeling Quarterly*, nº 9, p. 67–83.

Ravat F., Zhao Y. (2019). Data Lakes: Trends and Perspectives. In *International Conference on Database and Expert Systems Applications*, p. 304–313. Springer.

Russom P. (2017). Data lakes: Purposes, practices, patterns, and platforms. *TDWI White Paper*.

Sawadogo P. N., Scholly E., Favre C., Ferey E., Loudcher S., Darmont J. (2019). Metadata systems for data lakes: models and features. In *European conference on advances in databases and information systems*, p. 440–451.

Shvachko K., Kuang H., Radia S., Chansler R. *et al.* (2010). The hadoop distributed file system. In *Msst*, vol. 10, p. 1–10.

Simone C., Barile S., Calabrese M. (2018). Managing territory and its complexity: a decision-making model based on the viable system approach (vsa). *Land use policy*, vol. 72, p. 493–502.

Revealing the Conceptual Schemas of RDF Datasets - Extended Abstract

Subhi Issa, Pierre-Henri Paris, Fayçal Hamdi, Samira Si-said Cherfi

CEDRIC - Conservatoire National des Arts et Métiers
292 Rue Saint Martin, Paris, France
{subhi.issa,faycal.hamdi,samira.cherfi}@cnam.fr,pierre-henri.paris@upmc.fr

ABSTRACT. *This paper is an extended abstract of our work published at CAISE'20. The full paper is available at https://doi.org/10.1007/978-3-030-21290-2_20.*
RDF-based datasets, thanks to their semantic richness, variety and fine granularity, are increasingly used by both researchers and business communities. However, these datasets suffer a lack of completeness as the content evolves continuously and data contributors are loosely constrained by the vocabularies and schemes related to the data sources. In the context of the Web of Data and user-generated content, the conceptual schema is implicit. In fact, each data contributor has an implicit personal model that is not known by the other contributors. Consequently, revealing a meaningful conceptual schema is a challenging task that should take into account the data and the intended usage. In this paper, we propose a completeness-based approach for revealing conceptual schemas of RDF data. We combine quality evaluation and data mining approaches to find a conceptual schema for a dataset, this model meets user expectations regarding data completeness constraints. To achieve that, we propose LOD-CM; a web-based completeness demonstrator for linked datasets.

RÉSUMÉ. *Grâce à leur richesse sémantique, leur variété et leur granularité fine, les jeux de données fondés sur RDF sont de plus en plus utilisés par les chercheurs et les organisations. Cependant, ces jeux de données souffrent d'un manque de complétude en raison de l'évolution continue du contenu et le fait que les contributeurs ne sont pas tenus à respecter un vocabulaire et un schéma précis lors de la publication de leurs données. Dans cet article, nous proposons une approche fondée sur la complétude pour révéler les schémas conceptuels des données RDF. Nous combinons des approches d'évaluation de la qualité et de fouille de données pour trouver un schéma conceptuel pour un jeu de données, ce modèle répond aux attentes des utilisateurs en termes de complétude des données. Pour ce faire, nous proposons LOD-CM; un démonstrateur de complétude pour les jeux de données liés.*

KEYWORDS: *conceptual modeling, completeness, model quality, conceptual schema mining*

MOTS-CLÉS : *modélisation conceptuelle, complétude, qualité du modèle, extraction des schémas conceptuels*

1. Introduction

Data became a strategic asset in the information-driven world. One of the challenges for companies and researchers is to improve the display and understandability of the data they manage and use. However, exploiting and using data, like Linked Open Data (LOD), even if it is more and more accessible, is not an easy task, because data is often incomplete and lacks metadata. In this work we propose an approach for deriving conceptual schemas from existing data. This approach takes into account two facets; the universe of discourse represented by the data sources, and the user's needs represented by the user's decisions during the conceptual model construction. As the model should express the meaningful state of the considered dataset, we rely on a mining approach leading to taking into consideration the data model from a more frequent combination of properties. The relevancy of these properties is handled by integrating a completeness measurement solution that drives the identification of relevant properties. To meet user's requirements, we propose to construct the conceptual model on a *scratch card* manner where the user decides about the parts of the conceptual model to reveal according to her needs and constraints (Issa *et al.*, 2019).

2. Conceptual schemas derivation

The approach that we propose is an iterative process which infers a conceptual model complying the expected completeness. The process of inferring this model goes through four steps (cf. Figure 1): First, a subset of data that corresponds to the user's scope is extracted from the triple store. This subset is then transformed into transactions and a mining algorithm is applied. In our approach, for efficiency reasons, we chose the well-known FP-growth algorithm (Han *et al.*, 2004) (any other itemset mining algorithm could obviously be used). From the generated frequent itemsets, only a subset of these frequent itemsets, called "Maximal" (Grahne, Zhu, 2003), is captured. This choice is motivated by the fact that, on the one hand, we are interested in the *expression* of the frequent pattern and, on the other hand, the number of frequent patterns could be exponential when the transaction vector is very large. $\mathcal{MFP}$ is the set containing all maximal frequent patterns. Each pattern in $\mathcal{MFP}$ is then used to calculate the completeness of each transaction (regarding the presence or absence of the pattern) and, hence, the completeness of the whole dataset regarding this pattern. The final completeness value will be the average of all completeness value calculated for each $\mathcal{MFP}$ pattern. Finally, based on the completeness value and $\mathcal{MFP}$ that guarantees this value, a conceptual schema is generated. The classes, the attributes, and the relations of the model will be tagged with the completeness value. All these steps are integrated in an iterative process in such a way that the user could choose some parts in the generated model to refine. The data corresponding to the parts to refine is then extracted from the triple store, and the same steps are carried out to generate a new model.

Our prototype, called LOD-CM, implementing this process is available here: http://cedric.cnam.fr/lod-cm/.

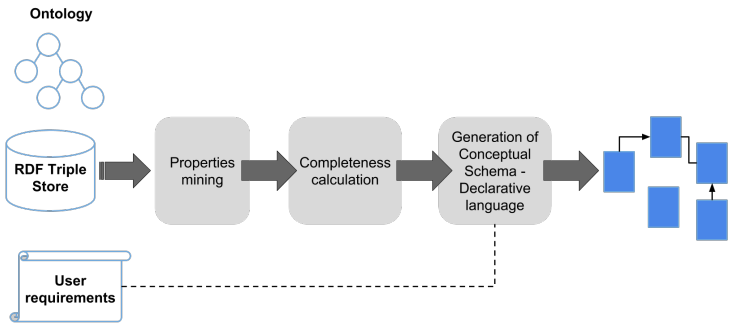


Figure 1. The LOD-CM Workflow.

3. Conclusion

In this paper, we presented an approach for revealing conceptual schemas from RDF data sources. Our approach is an iterative process that computes a plausible model from the data values. The inferred model takes into account the data and the user quality expectations. The result is a conceptual schema enriched by both completeness values as a relevancy indicator on the elements of the models, and existence constraints that inform about how often these elements co-exist or co-appear in the real data. In the future, we plan to investigate the role of conceptual modeling in an integration context where the universe of discourse is not only one data source but an integrated system upon several Linked Open Data. We plan to make more datasets available and allow the user to easily compare two conceptual schemas side by side (from two datasets). We believe that the ability to compare two conceptual schemas of two datasets side by side can help to choose the one that is best suited for its use.

References

Grahne G., Zhu J. (2003). Efficiently using prefix-trees in mining frequent itemsets. In B. Goethals, M. J. Zaki (Eds.), *FIMI '03, frequent itemset mining implementations, proceedings of the ICDM 2003 workshop on frequent itemset mining implementations, 19 december 2003, melbourne, florida, USA*, Vol. 90. CEUR-WS.org. Retrieved from http://ceur-ws.org/Vol-90/grahne.pdf

Han J., Pei J., Yin Y., Mao R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Discov.*, Vol. 8, No. 1, pp. 53–87. Retrieved from https://doi.org/10.1023/B:DAMI.0000005258.31418.83

Issa S., Paris P., Hamdi F., Cherfi S. S. (2019). Revealing the conceptual schemas of RDF datasets. In P. Giorgini, B. Weber (Eds.), *Advanced information systems engineering - 31st international conference, caise 2019, rome, italy, june 3-7, 2019, proceedings*, Vol. 11483, pp. 312–327. Springer. Retrieved from https://doi.org/10.1007/978-3-030-21290-2_20