Gestion de la qualité des données ouvertes liées – État des lieux et perspectives

Delphine Barrau*, Nathalie Barthélémy**, Zoubida Kedad***, Brigitte Laboisse****, Sylvaine Nugier♯, Virginie Thion♯♯

* Conix Consulting, France
Delphine.Barrau@conix.fr
** Gaz Réseau Distribution France (GRDF), France
Nathalie.Barthelemy@grdf.fr
*** Université Versailles Saint Quentin, DAVID Lab, France
Zoubida.Kedad@uvsq.fr
**** Base for Data Quality Systems (BDQS), France
blaboisse@bdqs.com
♯ Électricité de France (EDF) R&D, France
Sylvaine.Nugier@edf.fr
♯♯ Université Rennes 1, IRISA, France
Virginie.Thion@irisa.fr

Résumé. Sous l'impulsion de l'apparition de nouvelles technologies permettant la publication et l'exploitation de données ainsi que de contraintes réglementaires imposant à certaines entreprises et institutions d'ouvrir leur données, la publication de données liées est devenue un phénomène en pleine croissance. Cette immense ressource de données offre de grandes possibilités d'exploitation. Cependant, on constate un niveau très disparate de qualité des données publiées, rendant leur utilisation difficile, voir risquée. L'évaluation et la maîtrise de la qualité de ces données deviennent de ce fait un enjeu de premier plan. Dans cet article, nous dressons un état de l'art des approches méthodologiques et techniques de gestion de la qualité des données ouvertes liées proposées dans la littérature. Le périmètre couvert inclut les dimensions et métriques, les frameworks de gestion, les plateformes et outils associés, et enfin les cas d'étude de publication et d'utilisation des données ouvertes liées centrés sur la qualité de celles-ci. En nous fondant sur cet état de l'art, nous exhibons des problèmes ouverts et des perspectives de recherche du domaine.

1 Introduction

La publication de données ouvertes sur le web est un phénomène en pleine croissance. Il s'explique en partie par l'apparition de contraintes réglementaires et incitations gouvernementales (par exemple, incitation du secrétariat général pour la modernisation de l'action publique, de la loi sur le numérique) menant certaines entreprises, établissements publics et collectivités locales à exposer leurs données d'intérêt public. Au delà du souci de transparence portant cette démarche, un objectif espéré est la valorisation des données par l'exploitation de celles-ci par des tiers. Parmi les nombreuses données disponibles, on peut citer à titre d'exemple les données mises à disposition sur le portail interministériel data.gouv.fr [1] administré par la mission Etalab [2], et des données mises à disposition par des grandes entreprises dans les secteurs du transport comme la SNCF [3], ou de la production et commercialisation d'énergie comme EDF incluant les données de l'opérateur RTE [4]. De plus, le développement et la standardisation des technologies du web sémantique (par exemple outils liés au modèle RDF) amènent certaines communautés, telles que la communauté *Linked Open Data*, à mettre à disposition de gros volumes de données réelles, dans l'objectif, entre autres, de pouvoir expérimenter les nouvelles technologies développées. À ce jour, le nuage des données ouvertes liées (*Linked Open Data (LOD) cloud*) est constitué de plus d'un millier de jeux de données (voir LOD Catalog (2015)).

De nombreuses organisations ouvrent leurs données et les publient sur le Web. Il s'agit principalement d'administrations et de collectivités locales mais également de quelques entreprises faisant figure de précurseurs telles que le producteur et distributeur d'énergie Enel [5] ou le groupe JCDecaux [6]. Pour les administrations et organisations gouvernementales, cette ouverture est effectuée dans un souci de transparence et afin de garantir aux citoyens l'accès aux données publiques. Par la publication de leurs données, les administrations publiques visent à rendre leurs services plus accessibles à leurs usagers et à améliorer leur fonctionnement interne. De nombreuses initiatives sont menées pour promouvoir l'ouverture des données, dans le cadre notamment de la gouvernance en ligne (eGouvernment). Mais les entreprises ont également intérêt à ouvrir leurs données, au moins en partie : les données ouvertes peuvent leur permettre d'accélérer le développement d'un écosystème, d'agir sur leur image ou encore de retrouver la confiance de leurs clients. Dans certains domaines, les données ouvertes sont aussi bien utilisées par le grand public que par les professionnels des organisations qui les publient, c'est le cas des données du patrimoine. Un exemple d'initiative est la bibliothèque numérique Europeana, lancée par la Commission européenne et qui met en commun les ressources (livres, photos, matériel audio, etc.) des bibliothèques nationales des 27 états membres. L'un des objectifs de ce catalogue est de promouvoir les données ouvertes, et un grand sous-ensemble des données proposées est aujourd'hui disponible sous forme de données ouvertes liées utilisant les standards du Web sémantique. Les médias sont également à l'origine d'initiatives de publication de données. La BBC publie ainsi des données en lien avec les sujets abordés dans les émissions, qu'il s'agisse de lieux, de personnes ou d'organisations. Toutes ces initiatives montrent que l'intérêt des organisations pour la publication d'une partie de leurs données

1. http://www.data.gouv.fr
2. http://www.etalab.gouv.fr
3. https://data.sncf.com/
4. https://opendata.rte-france.com
5. http://data.enel.com
6. https://developer.jcdecaux.com

est certain, même si les usages et les applications qui les utiliseront sont encore largement à inventer.

Le potentiel d'exploitation des données ouvertes est bien reconnu par tous. Ainsi, de nombreux défis voient le jour, comme l'initiative Epidemium [7], un programme de recherche participatif visant à mieux comprendre le cancer grâce aux données ouvertes (Epidemium a repertorié et référencé plus de 21 000 jeux de données ouvertes), ou encore le challenge Open data *Des données au service de la ville intelligente* [8] visant à faire émerger des applications innovantes utilisant des données ouvertes, avec des données mises à disposition par les métropoles de Montpellier et de Rennes.

L'avènement des données ouvertes va même jusqu'à engendrer l'apparition de nouveaux concepts tels que l'*open business intelligence* (l'informatique décisionnelle exploitant des données ouvertes, voir Mazón et al. (2012)), et de nouvelles disciplines telles que la *science des données* [9] (l'extraction et l'exploitation de larges volumes de données en partie issues des données ouvertes), et le *journalisme des données* [10] (le journalisme dont une partie des sources est issue de la collecte puis de l'exploitation des données ouvertes), ainsi que des métiers en lien avec la *monétisation des données* ouvertes retravaillées. Dans de nombreux domaines, l'avènement des données ouvertes a ainsi occasionné l'apparition de nouveaux acteurs avec lesquels les acteurs historiques doivent composer. Ces nouveaux acteurs ont accès à une donnée gratuite et peuvent la confronter à de multiples sources pour la qualifier, la valider ou offrir des solutions plus fiables, plus exhaustives et beaucoup moins onéreuses. Par exemple, avec l'ouverture d'une partie des données de l'INSEE à compter de 2017, incluant le célèbre repertoire SIRENE (enregistrant l'état civil et les établissements de toutes les entreprises françaises), le monde du BtoB se prépare à une petite révolution. La tendance qui se dessine va dans le sens de l'émergence d'un métier d'*intermédiaire* de la donnée, en charge du travail de mise en forme, validation, croisement des données dans l'objectif de pouvoir offrir à un utilisateur des données prêtes à l'emploi. Du point de vue des utilisateurs des données ouvertes, une problématique centrale est celle de la qualité des données publiées, qui conditionne la qualité des résultats de toutes les applications qui les exploitent et permettent d'en extraire des connaissances nouvelles.

Beaucoup de données publiées souffrent cependant d'un manque de qualité (Zaveri et al. (2016)). L'ouverture des données permet d'ailleurs, par croisement de sources entre elles, de prendre conscience d'incohérences qui n'auraient pas pu être détectées auparavant. Dans ce contexte, grande est la pression pesant sur les producteurs dont les données doivent respecter un certain niveau de qualité au risque de mettre en péril leur crédibilité, et difficile est la tâche des consommateurs de données confrontés à un grand volume de données très hétérogènes présentant des niveaux de qualité divers. On assiste donc depuis quelques années à un intérêt de plus en plus marqué des entreprises et des chercheurs pour le sujet de la gestion de la qualité des données ouvertes, pouvant être liées (publiées sous cette forme ou transformées).

Les données ouvertes, éventuellement liées, présentent certaines caractéristiques inhérentes : elles sont ouvertes à tous, volumineuses, issues de diverses sources de publication, de points de vue et de périmètres différents, et sont éventuellement liées les unes aux autres. Nombre de travaux relatifs à la gestion de la qualité des données, développés dans le cadre des données non

7. `http://www.epidemium.cc/`
8. `http://www.entreprendre-montpellier.com/challenges-big-data`
9. *Data science*.
10. *Data journalism*.

ouvertes et/ou non liées, ne sont plus applicables ou nécessitent d'être adaptés pour pouvoir être utilisés dans ce cadre. D'autre part, l'accès ouvert, l'homogénéité syntaxique visée par le modèle RDF et les informations enrichissant la sémantique de données (en particulier, les ontologies associées et les liens établis entre sources de données) peuvent faciliter l'intégration et l'enrichissement des données, offrant de nouvelles perspectives d'exploitation de celles-ci. Enfin, si la publication des données ouvre de belles perspectives, elle pose aussi de nouveaux problèmes liés au manque de connaissance ou de contrôle de l'utilisation qui peut être faite des données.

Dans cet article, nous dressons un état de l'art des approches méthodologiques et techniques de gestion de la qualité des données ouvertes liées proposées dans la littérature. Cette étude est présentée dans la section 2. Nous présentons ensuite dans la section 3 des cas d'étude et retours d'expérience concernant l'exploitation et la publication de données ouvertes. Ces travaux nous permettent d'identifier des verrous méthodologiques et techniques restant à lever, présentés sous forme de perspectives de recherche dans la section 4. L'un de nos objectifs est de présenter ces travaux sous l'angle des évolutions induites, en termes de gestion de la qualité des données, par les caractéristiques d'ouverture et d'interconnexion des données, par rapport au cadre classique. Par cadre classique, nous entendons ici celui de la gestion de donnés multi-sources hétérogènes mais non ouvertes et non liées. Pour finir, nous dressons une conclusion de notre étude en section 5.

2 Approches de la littérature

Certains travaux ont abordé le problème de la gestion de la qualité des données ouvertes liées. Les contributions peuvent être classées en trois catégories : les contributions relevant de la définition de la qualité, celles relevant de l'amélioration de la qualité des données, et les outils ou plateformes associés à ces contributions.

2.1 Définition de la qualité

La qualité des données est définie par « l'adéquation des données à l'usage qui en est fait ». Il s'agit de la notion de *fitness for use* que l'on retrouve dans Wang et Strong (1996), Strong et al. (1997) et Batini et al. (2009). Cette définition générale est à instancier en fonction du contexte opérationnel d'utilisation des données. Guidé par une méthodologie, il est question de définir en contexte les indicateurs (dimensions et métriques qualité) d'intérêt permettant de mesurer l'adéquation des données aux utilisations qui en sont faites. Cette instantiation de definition de la qualité des données est un problème central dans le processus global de gestion de la qualité des données.

La littérature fournit un grand nombre de méthodologies et métriques issues du cadre classique (voir Batini et Scannapieco (2006); Batini et al. (2009)). Étant donnée la définition de la qualité des données (adéquation des données à l'usage qui en est fait), les méthodes classiques de définition de la qualité des données sont guidées par l'utilisation faite des données. Mais dans le cadre de données ouvertes liées, toutes les utilisations faites des données ne sont pas connues au moment de la publication des données. Il en résulte une partielle inadéquation des méthodes classiques de définition de la qualité des données au contexte des données ouvertes.

Il apparaît clairement que les enjeux sous-jacents à l'utilisation des données ouvertes éventuellement liées évoluent par rapport au cadre classique, donc les dimensions, métriques, ontologies, vocabulaires et méthodologies de gestion de la qualité des données doivent être définis ou adaptés pour le contexte des données ouvertes liées.

2.1.1 Dimensions et métriques de la qualité

Zaveri et al. (2016) dressent un état de l'art des métriques et dimensions proposées dans la littérature pour mesurer la qualité des données ouvertes liées. Le tableau 1 rappelle les définitions des principales dimensions qualité de la littérature. Cette étude traite de dimensions classiques de la qualité des données telles que la *complétude*, la *pertinence*, la *fraîcheur*, l'*exactitude*, la *consistance*, dont l'interprétation générale change assez peu du cadre classique de l'intégration d'information de données multi-sources. D'autre dimensions comme la *confiance* ou la *traçabilité* sont abordées sous de nouvelles perspectives. Enfin, de nouvelles dimensions telles que l'*interconnexion* sont apparues, spécifiques au cas des données ouvertes liées.

Nous présentons ci-dessous quelques travaux représentatifs des dimensions et métriques qualité associées aux données ouvertes liées, ces travaux étant présentés sous l'angle des évolutions induites par les caractéristiques d'ouverture et d'interconnexion des données.

N'importe qui peut publier des données. Des données ouvertes liées peuvent être publiées par « n'importe qui ». La dimension de *confiance* des données et des sources prend donc ici toute son importance. La notion de confiance est très liée à la celle de *traçabilité* des données, puisque la confiance est souvent établie au regard de l'information de la provenance des données. La provenance ne concerne pas seulement l'identification du producteur d'une donnée mais concerne également les informations relatives au processus de production et de publication de la donnée. Hartig et Zhao (2009) ont constaté un manque d'information relevant de la provenance des données ouvertes liées. Cette situation est en partie explicable par le manque d'outils méthodologiques et techniques permettant d'exploiter ce type d'information. Certains travaux ont donc porté sur l'enrichissement des données par des informations de provenance (voir section 2.1.2). Ces informations peuvent ensuite être exploitées afin d'évaluer la qualité des données récoltées. Dans les travaux de Hartig et Zhao (2009), les informations de provenance sont utilisées pour enrichir le calcul de métriques qualité liées à la fraîcheur des données. Dans les travaux de Gil et Ratnakar (2002), une méthode d'évaluation du niveau de confiance des données est définie en fonction d'annotations collaboratives effectuées sur celles-ci.

Les données reflètent la vision du fournisseur. Les données mises à disposition par un fournisseur, sont présentées sous le prisme de sa vision. Leur couverture, leur représentation, le vocabulaire utilisé, leur niveau de granularité reflètent ce point de vue. Les données de sources différentes peuvent donc être qualifiées d'hétérogènes, sans pour autant associer une connotation négative à ce terme puisque les données ouvertes présentent intrinsèquement cette caractéristique. Détecter des données redondantes ou contradictoires entre diverses sources, ou chercher à évaluer leur complétude, s'avèrent donc être des tâches complexes voir inadaptées à ce cadre. Par ailleurs, des données peuvent être considérées de bonne qualité du point de vue du fournisseur, et ce de façon complètement justifiée de ce point de vue, mais ne correspondre

Dimension	Définition
Exactitude	L'exactitude syntaxique définit à quel point les données sont conformes à une règle de format (par exemple, une donnée est-elle bien un numéro de téléphone ?). L'exactitude sémantique définit à quel point les données sont conformes à la réalité décrite (par exemple, le numéro de téléphone est-il bien celui de l'entité décrite ?).
Vérifiabilité	Définit à quel point les données peuvent être contrôlées, certifiées ou garanties par contrat.
Traçabilité	Définit à quel point les données portent l'information de leur *provenance* (par exemple source d'origine, processus de transformation subi avant publication, identification du producteur).
Confiance	Définit à quel point les données et leur producteur sont fiables.
Pertinence	Définit à quel point les données apportent une valeur ajoutée dans leur utilisation.
Utilisabilité	Accessibilité : Définit à quel point les données sont disponibles, récupérables. Compréhensibilité : Définit à quel point les données sont compréhensibles, incluant par exemple l'éventuelle présence d'un support et d'une documentation, ou mesurant la versatilité des données. *Licensing* : Présence ou non d'une license indiquant quelle ré-utilisation peut être faite des données.
Visibilité	Définit à quel point les données sont localisables par les utilisateurs.
Fraîcheur	Définit à quel point les données sont suffisamment récentes.
Complétude	Définit le niveau de couverture avec lequel le phénomène observé est représenté dans l'assemblage des données. Se décline en complétude de la population, et des informations présentes au niveau des individus.
Cohérence	Définit à quel point les données satisfont un ensemble de contraintes syntaxiques et sémantiques.
Unicité	Définit à quel point les données évitent les redondances (des données sont redondantes si elles décrivent un même objet du monde réel).
Consistance	Définit à quel point les données évitent les contradictions et incohérences, peut concerner la détection de données redondantes (décrivant un même objet du monde réel) mais contradictoires ou la vérification de règles de cohérence métier.
Sécurité de l'accès	Définit à quel point l'accès aux données est contrôlé (peut être vue comme une sous-dimensions de celle d'*accessibilité*).
Confidentialité	Définit à quel point la confidentialité des informations personnelles est préservée.
Interconnexion	Définit à quel point les données sont riches et précises en termes de lien vers des sources externes complémentaires.

TAB. 1 – *Définition des principales dimensions qualité*

qu'en partie aux besoins du consommateur. Des données de bonne qualité du point de vue du producteur ne sont donc pas nécessairement des données de bonne qualité pour tous les consommateurs.

Les données sont liées. La dimension d'*interconnexion* apparaît dans le contexte spécifique des données liées. Elle regroupe des métriques permettant de qualifier la richesse des liens entre (jeux de) données. Des métriques permettant de qualifier cette dimension ont donc été proposées. Par exemple, Guéret et al. (2012) proposent des métriques inspirées du domaine de l'analyse structurelle de réseaux afin d'évaluer la qualité des mappings entre données liées.

De nouvelles techniques sont introduites pour le traitement des données liées, de façon à exploiter les liens entre données. Par exemple, Ruckhaus et al. (2014) proposent une méthode fondée sur des techniques d'analyse statistique pour étudier la qualité des données et des liens entre données. L'objectif est d'exhiber des données pour lesquelles on suspecte des problèmes de qualité, limités aux dimensions de complétude, unicité ou consistance. Cette méthode tire en partie sa richesse de l'exploitation et de l'analyse des liens.

L'utilisation faite des données n'est pas connue *a priori*. La dimension de *pertinence* qualifie l'utilité (la valeur ajoutée) des donnés pour atteindre l'objectif fonctionnel. Comme toutes les dimensions qualité, la dimension de *pertinence* des données est dépendante de l'utilisation faite de celles-ci, mais la définition de cette dimension en particulier est intrinsèquement et complètement dépendante du contexte et du cas d'étude traité, ce qui laisse peu de place à la définition de métriques génériques permettant de la qualifier. Ceci explique la rareté des publications scientifiques universitaires visant à proposer des métriques participant à la définition de cette dimension. La *pertinence* des données se révèle pourtant être d'un grand intérêt pour la communauté industrielle car elle constitue un critère discriminant dans le choix des sources de données ouvertes à utiliser. Cette notion est très liée à la notion de *valeur*, étudiée par Chignard et Benyayer (2015) et GDEF (2015).

Comme l'utilisation faite des données n'est pas connue, une licence d'utilisation, indiquant quelle ré-utilisation peut être faite des données, doit être associée aux données. La dimension de *licensing* regroupe des métriques consistant à vérifier la présence de cette information, comme proposé par Hogan et al. (2012).

Enfin, il est intéressant de constater qu'une nouvelle métrique permettant de qualifier en partie la dimension de *pertinence* des données pourrait être le nombre d'utilisations faites des données.

Les données sont ouvertes à tous. Il est possible de tirer parti de l'aspect ouvert des données pour appliquer des techniques d'évaluation de leur qualité reposant sur une évaluation collaborative. Dans les travaux de Gil et Ratnakar (2002), le niveau de confiance des données est évalué de façon collaborative à partir d'annotations apposées sur les données par leurs utilisateurs. D'autres travaux proposent des méthodes d'évaluation reposant sur un processus de *crowdsourcing*. Le *crowdsourcing* consiste à proposer à un ensemble d'individus d'assumer volontairement des tâches, afin d'atteindre un objectif global plus complexe. Les contributeurs sont sollicités et contribuent en ligne, éventuellement contre rémunération, en ayant conscience ou non de l'objectif global auquel ils participent. Dans les méthodes proposées par Kontokostas et al., Zaveri et al. (2013) et Acosta et al. (2013), l'objectif global à atteindre est l'évaluation

de la qualité d'un jeu de données, les contributeurs ayant pour tâche de relever des problèmes détectés dans les données. Ces travaux proposent également une mise en correspondance de problèmes de qualité avec des métriques et dimensions, dans l'objectif de simplification de la tâche attribuée aux contributeurs pour lesquels il est plus naturel de penser en termes de problèmes rencontrés qu'en termes de métriques et dimensions qualité. Ce travail est dépendant du contexte et doit donc être mené pour chaque contexte d'évaluation. Enfin, des techniques de corroboration de sources comme proposé par Galland et al. (2010) peuvent également être appliquées.

L'accès aux données se fait via le web. Sans une qualité suffisante de l'*accessibilité* aux données, celles-ci ne peuvent plus être considérées comme ouvertes, comme l'illustre la récente pétition lancée par *Citymapper* demandant une meilleure accessibilité des données de la RATP. Des métriques permettant de qualifier l'*accessibilité*, peu dépendantes du domaine d'application, ont par exemple été proposées par Hogan et al. (2010). On peut citer à titre d'exemple la mesure du nombre de liens brisés, la disponibilité et le temps de réponse du serveur hébergeant les données ou encore la mise en œuvre d'une publication sécurisée via le protocole SSL. Mais l'accessibilité ne concerne pas que les aspects techniques. Dans le cadre des données ouvertes, se pose également le problème de l'*agilité* (à publier et à corriger) des fournisseurs de données, et leur capacité à assurer un flux régulier de données exploitables. La survie de nombreuses entreprises émergentes dont le métier et les outils développés reposent sur l'exploitation de données ouvertes dépend du niveau d'accessibilité des données publiées.

Les données ouvertes peuvent représenter de gros volumes à traiter. Les données ouvertes liées constituent de gros volumes et sont susceptibles d'évoluer de façon rapide. Des travaux se concentrent donc sur la définition de métriques qualité assurant un passage à l'échelle de leur mesure. Debattista et al. (2015) proposent des métriques qualité dont la valeur est calculée de façon probabiliste (donc approximative) et efficace.

Le tableau 2 synthétise notre vision de l'adaptation des dimensions qualité au monde des données ouvertes liées.

Les dimensions et métriques ont vocation à être inscrites dans des concepts méthodologiques plus complexes que sont les méta-modèles, ontologies et vocabulaire qualité, présentés dans la section suivante.

2.1.2 Méta-modèles, ontologies et vocabulaire

Les méta-modèles ou vocabulaires décrivant les concepts liés à la qualité des données du web sémantique ont fait l'objet de quelques propositions, comme le modèle conceptuel pour la gestion de la qualité présenté dans Fürber et Hepp (2011b). L'ontologie proposée est utilisée pour exprimer des règles qui régissent la qualité des données ainsi que les règles de nettoyage de ces données. Elle fournit également une classification des problèmes de qualité des données et décrit le calcul de scores de qualité pour les données des sources.

Fürber et Hepp (2011a) ont proposé SWIQA, une méthode d'évaluation de la qualité des données du web utilisant une classification des problèmes de qualité et définissant des mé-

Dimension	Adéquation aux LOD
Exactitude	A
Vérifiabilité	A+
Traçabilité	A+
Confiance	A+
Pertinence	A
Utilisabilité	A
Visibilité	A+
Fraîcheur	A
Complétude	D
Cohérence	D
Unicité	D
Consistance	D
Sécurité de l'accès	N
Confidentialité	A+
Interconnexion	Dimension apparue dans le contexte des LOD.

Légende :
A : Applicable avec adaptation naturelle au cadre des LOD.
A+ : Applicable, plus important dans le domaine des LOD que dans le cadre classique.
N : Non applicable.
D : Difficilement applicable.

TAB. 2 – *Adaptation des dimensions qualité au monde des données ouvertes liées*

triques permettant d'attribuer un score à chacun. Cette méthode est implantée dans un outil du même nom.

Dans Hartig et Zhao (2010), un vocabulaire est proposé permettant l'enrichissement de données RDF par des informations caractérisant la provenance des données. Ce vocabulaire a été implanté dans plusieurs outils de publication de données afin de permettre d'automatiser l'enrichissement des sources.

Une méthodologie pour évaluer la qualité des sources de données liées est proposée dans Acosta et al. (2013). Elle comporte deux phases distinctes, la première porte sur la détection de problèmes de qualité ; une taxonomie décrivant ces problèmes est proposée. La seconde phase porte sur l'évaluation proprement dite de la qualité d'un ensemble de sources de données en utilisant la technique du *crowdsourcing*.

Enfin, Debattista et al. (2014) proposent une ontologie extensible permettant d'adjoindre à des jeux de données RDF les résultats de l'évaluation d'un ensemble de métriques de qualité. Cette ontologie est également utilisée pour une navigation guidée par la qualité dans la source de données.

2.1.3 Méthodologies et recommandations

De nombreuses méthodologies de gestion de la qualité des données ont été définies dans un cadre classique. Elles décrivent comment définir, mesurer, améliorer et suivre (donc comment gérer) la qualité de données d'intérêt en fonction de l'utilisation faite de celles-ci. Nous pouvons citer à titre d'exemple la méthode GQM de définition de la qualité des donnés de Basili et al. (1994) et le cycle DMAIC de gestion de la qualité de De Feo et al. (2005). Nous renvoyons le lecteur intéressé par ces méthodologies vers les travaux de Batini et al. (2009) qui en proposent un état de l'art.

Du point de vue du consommateur de données, ces méthodes restent pertinentes pour l'exploitation de données ouvertes liées, via l'adaptation des métriques, dimensions et méta-modèles considérés (voir section 2.1.1 et section 2.1.2), ainsi que des méthodes d'amélioration des données utilisables (voir section 2.2).

En revanche, le fait que toutes les utilisations faites des données ne soient pas connues au moment de leur publication rend le contexte plus délicat pour le producteur de données. Pour illustrer cela, plaçons nous du point de vue d'un fournisseur qui publie des données concernant une étude menée sur les centres de tatouage en Europe comportant entre autres pour chaque centre la ville associée et une adresse mail de contact si celle-ci est connue. Ces données ont été créées avec le plus grand soin à partir de données recueillies l'année dernière, et mises à disposition sous forme d'un fichier pdf comportant un dictionnaire de données. Du point de vue du fournisseur, ces données sont de bonne qualité. Elles ont été vérifiées (leur niveau d'exactitude est bon), elle sont complètes, précises, documentées et fraîches. Ces données sont de qualité tout à fait adéquates pour un consommateur 1 qui désire connaître le nombre de centres de tatouage en France l'année dernière. Mais elles ne sont pas de qualité suffisante pour un consommateur 2 qui désirerait créer un fichier de prospects car ces données ne sont ni suffisamment complètes (l'adresse n'est pas complète et le mail n'est pas toujours présent), ni suffisamment fraîches (des données du mois dernier seraient nécessaires), ni dans un format adéquat (il est compliqué pour le consommateur 2 d'extraire des données fournies au format PDF). La définition de la qualité des données du point de vue du fournisseur ne correspond donc pas à la définition de la qualité des données pour tous les consommateurs. Comme dans

la cadre classique, le fournisseur a pourtant tout intérêt à connaître les besoins qualité liés à ses données s'il désire qu'elles soient exploitées au mieux. En admettant qu'un fournisseur se soucie de l'utilisabilité de ses données, comment pourrait-il avoir connaissance des besoins qualité liés à ses données puisqu'il n'en connaît pas les utilisations faites ? Ce contexte rend inapplicables les méthodologies classiques de définition de la qualité dont l'objectif est la définition de la qualité à partir d'un besoin fonctionnel. Tout processus de gestion de la qualité des données repose pourtant sur une première étape de définition de la qualité. Quelques contributions de la littérature tentent d'apporter une réponse à ce problème.

Behkamal et al. (2014) ont mené une étude empirique visant à évaluer la pertinence de quelques métriques sur les dimensions les plus utilisées de qualité des données (exactitude, unicité, consistance, complétude). Une telle étude ne peut être menée que sur une quantité réduite de métriques, parmi les plus communes, et la validation complète de ce type d'approche nécessite des interventions humaines d'experts métier. La mise en correspondance avec des cas d'utilisation réels n'est pas étudiée.

Il existe quelques méthodologies de haut niveau proposées dans l'objectif de guider un projet de publication de données. Le W3C Working Group (2014) et le W3C Government Linked Data Working Group (2011) ont par exemple proposé des méthodologies assez générales pour la publication de données gouvernementales. Ces méthodologies sont composées de la description des phases composant un processus de publication des données, enrichies de bonnes pratiques à mettre en œuvre pour chacune de ces phases. L'objectif de ces méthodologies est d'améliorer la qualité des données publiées. En complément à ces méthodologies, Hogan et al. (2010) ont mené une étude consistant à analyser la qualité d'un ensemble de données ouvertes liées dans l'objectif de découvrir des problèmes de qualité. À partir des constats dressés, les auteurs proposent de bonnes pratiques à mettre en œuvre au moment de la publication ou de l'exploitation des données afin d'éviter l'apparition des problèmes rencontrés lors de l'étude. Cette étude montre que les bonnes pratiques introduites par le W3C ne sont pas toujours suivies, une partie des recommandations à destination des producteurs de données rejoignent donc celles du W3C Working Group (2014).

Parmi ces bonnes pratiques de la littérature, on peut citer l'importance d'une phase de modélisation soignée des données à publier. Les données à publier doivent être choisies en cherchant à anticiper les utilisations qui pourraient être faites non seulement dans l'objectif de mieux cibler les besoins des futurs utilisateurs, mais également pour maîtriser les risques engendrés par un usage des données (ce sujet est développé dans la section 3). Il convient également de chercher à réduire les effets engendrés par une éventuelle trop grande *versatilité* des données qui rendrait les données trop rapidement *obsolètes*. Une trop grande versatilité dégrade la *compréhensibilité* des données. Ces bonnes pratiques constituent des guides très généraux qui doivent être adaptés à chaque contexte, ce qui reste une tâche extrêmement compliquée pour les producteurs de données. D'autres bonnes pratiques, plus pragmatiques, sont associées à l'amélioration de la *représentation* des données. On peut citer l'utilisation des standards RDF et SPARQL, leur respect strict étant contrôlé par les outils de restructuration, normalisation et validation syntaxiques [11] (impactant les dimensions d'*exactitude syntaxique* et d'*utilisabilité*) ; l'utilisation poussée des URI pour nommer les objets modélisés et les enrichir (impactant la dimension d'*interconnexion*) ; l'utilisation de clefs naturelles (impactant les dimensions de *compréhensibilité* et d'*utilisabilité*) ; l'utilisation et le maintien d'ontologies

11. Par exemple, les outils recensés par W3C (b).

associées aux données (impactant les dimensions de *complétude*, *traçabilité* et *consistance*) ; l'association systématique d'information de provenance des données (*fiabilité*) ; et la mise à disposition des données via une API d'accès fiable et standardisée telle que l'API RESTful ou le SPARQL endpoint (impactant la dimension d'*accessibilité*).

2.2 Amélioration de la qualité

En plus des propositions visant à définir les dimensions qualité et les métriques associées et à proposer des méthodologies d'évaluation de la qualité, un certain nombre de propositions ont été faites pour traiter le problème de l'amélioration de la qualité des données ouvertes et liées sous différents angles. Ainsi, une approche permettant de filtrer l'information selon sa qualité est présentée dans Bizer et Cyganiak (2009). Elle permet de fournir à l'utilisateur un support pour filtrer l'information ayant une qualité élevée parmi l'ensemble des données disponibles. Ce filtrage consiste à exécuter un ensemble de métriques de qualité, puis à utiliser une fonction de décision utilisant l'agrégation des scores obtenus.

Mendes et al. (2012) proposent *Sieve*, un cadre pour la fusion et l'évaluation de la qualité des données liées. Intégré dans la plate-forme d'intégration *LDIF*, introduite par Schultz et al. (2011), il permet entre autres la résolution d'entités, et attribue une même URI à plusieurs identifiants distincts représentant le même objet. La fusion de données conflictuelles provenant de plusieurs sources se fait en utilisant le résultat de l'évaluation de la qualité des données.

Alors que la plupart des approches s'intéressent à l'amélioration de la qualité des données elles-mêmes, certaines, comme celle présentée dans Dimou et al. (2015), proposent d'améliorer la qualité des mappings de sources de données RDF générées à partir de sources structurées ou semi-structurées telle que CSV, XML ou Json. Une approche guidée par les tests est proposées, où les mappings sont redéfinis et affinés en fonction des résultats de l'évaluation de leur qualité.

Kontokostas et al. (2014) proposent une méthodologie, inspirée de techniques de développement logiciel guidé par les tests, permettant de détecter les problèmes de qualité dans des données ouvertes liées. Des cas de tests détectant des problèmes de qualité des données sont définis, sur la base de l'instanciation semi-automatisée d'un ensemble de patrons pré-définis (exprimés en langage SPARQL), permettant la mise ne œuvre d'une procédure d'amélioration. Une expérimentation d'envergure est présentée, couvrant plusieurs jeux de données ouvertes.

Enfin, certaines approches se sont intéressées à l'amélioration de la complétude des informations sur le type des ressources représentées dans un jeu de données, dans le cas où ces informations sont partiellement définies. C'est le cas de l'approche présentée dans Paulheim et Bizer (2014), qui permet, en utilisant la distribution des types, d'ajouter les définitions de types manquantes ou d'identifier celles qui peuvent être erronées. Arenas et al. (2014) utilise un solveur pour spécialiser les classes en divisant leurs entités selon la proximité des propriétés qui les caractérisent et la co-occurrence de ces propriétés.

Certains des travaux présentés ci-dessus ont donné lieu à l'implantation d'outils, que nous présentons dans la section suivante.

2.3 Plateformes, outils support

De nombreux outils ont été proposés par les communautés de recherche et industrielle. Nous en donnons ci-dessous un aperçu, en classant les outils selon trois grandes catégories : les outils supports dédiés à l'évaluation ou à l'amélioration de la qualité des données, les outils supports à la publication de la qualité des données, et enfin les plateformes.

Outils supports à l'évaluation ou à l'amélioration de la qualité des données

De nombreux outils dédiés à l'évaluation et/ou l'amélioration (reposant sur une évaluation préalable) de la qualité des données ouvertes ont été développés.

On peut considérer les outils permettant de convertir des données de divers formats sources au format cible RDF comme étant des outils de restructuration de données visant à améliorer leur qualité. Une liste de ces outils est maintenue par le W3C (a). D'autres outils permettent de valider le format de données liées, tels quel RDF Validator, RDF Alerts, ou RDFa Developer. Ils effectuent essentiellement des vérifications syntaxiques de format ou d'adéquation à des concepts introduits dans une ontologie associée aux données. Une liste de ces outils est également maintenue par le W3C (b).

Nombre d'outils sont également proposés par la communauté de recherche et reposent sur les travaux présentés dans les sections précédentes (sections 2.1.1 et 2.2). L'outil *LINK-QA* repose sur le framework introduit par Guéret et al. (2012). Il automatise l'évaluation de métriques qualité sur des données liées, sur la dimension d'interconnexion des données. LINK-QA est développé en JAVA, il utilise l'outil Jena pour interagir avec les données RDF, et l'outil Any23 pour déréférencer les données. L'outil *LiQuate* implante la méthode proposée par Ruckhaus et al. (2014) fondée sur des techniques d'analyse statistique pour analyser la qualité des données ouvertes liées. L'outil *LUZZU* repose sur la méthode proposée par Debattista et al. (2015). Il permet l'évaluation d'une trentaine de métriques qualité. L'évaluation de ces métriques est voulue probabiliste et efficace de façon à pouvoir gérer de grands volumes de données. L'outil *GovWILD* proposé par Böhm et al. (2012) est dédié à l'intégration et au nettoyage de données gouvernementales ouvertes. Cet outil utilise les fonctions de croyance pour évaluer le niveau de confiance des données en cas de présence de données contradictoires. L'outil *RDF Unit* met en œuvre la méthode proposée par Kontokostas et al. (2014) consistant à vérifier la validité des données par rapport à des contraintes spécifiées sous la forme de patrons exprimés dans un langage dérivé de SPARQL. L'outil *ODCleanStore (ODCS)* proposé par Knap et al. (2012) permet l'agrégation de données ouvertes liées avec calcul du niveau de qualité des données agrégées. Le calcul est effectué au moment de l'évaluation des requêtes, intégrant une phase de résolution de conflit en cas de présence de données contradictoires. L'outil *Sieve* repose sur la méthodologie introduite par Mendes et al. (2012). Il permet de fusionner des données ouvertes liées en se fondant sur leur niveau de qualité. *LDIF* (Linked Data Integration Framework) proposé par Schultz et al. (2011) permet de créer des jeux de données ouvertes liées, porteuses d'informations de provenance, à partir de l'intégration de données hétérogènes. LDIF inclut également l'outil *Sieve* présenté ci-dessus. L'outil *WIQA* adossé au framework du même nom permet à un utilisateur de définir des patrons sous forme de graphes *nommés* (introduits dans Carroll et al. (2005)), à partir desquels les données peuvent être filtrées. L'outil *SWIQA* proposé par Fürber et Hepp (2011a) repose sur le framework du même nom permettant d'évaluer la qualité de données ouvertes liées par vérification de règles qualité définies sous la forme de patrons.

Des outils reposent sur la mise en œuvre des méthodes collaboratives d'évaluation de la qualité des données. L'outil *TripleCheckMate* développé par Kontokostas et al. permet l'évaluation de la qualité de données ouvertes liées par *crowdsourcing*. L'outil permet à un utilisateur participant à l'évaluation, de visualiser une source (par exemple tirée aléatoirement) et d'indiquer, au niveau le plus fin du triplet RDF, la présence de problèmes qualité parmi une taxonomie pré-définie de problèmes. La mise en correspondance des problèmes qualité décelés avec des métriques qualité et l'analyse des résultats obtenus par agrégation des réponses des contributeurs permettent d'associer un niveau de qualité aux données. L'outil *TRELLIS* apporte un support technique à une méthode d'évaluation de la qualité des données reposant sur une annotation collaborative des données proposée par Gil et Ratnakar (2002).

Les auteurs de Debattista et al. (2016) ont mené une étude comparative poussée des fonctionnalités offertes par quelques uns des outils listés ci-dessus, vers laquelle nous renvoyons de lecteur pour plus de détails.

Outils supports à la publication de la qualité des données

Les outils support dédiés à la *publication* des données ouvertes telles que CKAN (2016), Callimachus (2016) ou OpenDataSoft (2016) sont encore peu nombreux. Il s'agit de plateformes intégrant des modules d'enrichissement des données par des meta-données descriptives et des modules de navigation ou de visualisation des données, dans un objectif de valorisation des données publiées (par exemple, si les données publiées sont des données provenant de multiples capteurs embarqués alors l'outil OpenDataSoft met à disposition les données brutes et propose en complément une visualisation de la répartition des données sur une carte géographique). L'outil Aggrego (2016), développé par la société SemSoft, propose également un module de publication dans le sens où il met en œuvre un mécanisme de médiation qui permet, sur la base d'une ontologie, de construire dynamiquement une API d'accès à des données à consommer ou à publier.

Les outils existants en support au processus de publication de données sont des solutions apportant une aide essentiellement technologique, non méthodologique, à la publication de données. On peut considérer que ces outils proposent des fonctionnalités de gestion de la qualité des données dans le sens où chacun d'eux permet l'attachement de meta-données d'intérêt aux données. En revanche, seuls quelques outils affiche une gestion explicite de la qualité des données, sous forme de calcul et d'exploitation de métriques qualité.

Plateformes

Certains des outils supports présentés précédemment ont éventuellement vocation à être intégrés, par exemple sous forme de modules, à des plateformes de gestion d'une partie du cycle de vie des données ouvertes. Une telle plateforme cherche à mettre en œuvre une chaîne allant de la récupération des données (internes et externes), en passant par la détection et la résolution des problèmes de qualité, l'intégration des données (pouvant être guidée par la qualité des données), jusqu'à une éventuelle la publication et visualisation des données (pouvant intégrer la gestion de préférences utilisateurs). La plateforme *LOD2*, proposée par Auer et al. (2012), en est un exemple. Elle permet l'intégration d'outils tiers dans l'objectif de la gestion de données ouvertes liées sur tout leur cycle de vie. Un autre objectif de cette plateforme est d'offrir suffisamment d'adaptabilité pour pouvoir intégrer des modules externes de traitement

Outil	Publication	Consommation	Gestion qualité (*)	Type (**)
LINK-QA		×	×	L
LiQuate		×	×	L
LUZZU		×	×	L
GovWILD		×	×	P
RDF Unit		×	×	L
ODCleanStore		×	×	L
Sieve		×	×	L
WIQA		×	×	L
TripleCheckMate		×	×	L
TRELLIS		×	×	L
LDIF		×	×	L
Aggrego	×	×	×	C
LOD2	×	×	×	P
CKAN	×			L
Callimachus	×			L/C
OpenDataSoft	×			C

(*) On entend ici une gestion explicite s'appuyant sur une méthodologie.
(**) Logiciel Open source ou libre (L) / logiciel Commercial (C) / Prototype (P)

TAB. 3 – *Tableau synthétique des outils discutés*

(l'outil *Sieve* fait actuellement partie des modules intégrés par défaut à LOD2). Il existe également des outils commerciaux proposant des plateformes complètes, comme l'outil *Aggrego*.

Le tableau 3 propose une vue résumée des systèmes discutés ci-dessus.

3 Cas d'étude et retour d'expérience

Nous distinguons ici deux types de cas d'étude centrés sur la gestion de la qualité des données ouvertes : ceux relevant de l'utilisation des données et ceux relevant de leur publication.

3.1 Utilisation de jeux de données guidée par la qualité

Des travaux de la littérature concernent des études intégrant une évaluation de la qualité des données de la plate-forme collaborative Open Street Map [12]. Open Street Map attire une grande attention car les données géographiques qui y sont publiées constituent des données ayant une grande utilité dans nombre de domaines. Nous pouvons citer à titre d'exemple les travaux de Arsanjani et al. (2015), Fan et al. (2014), Forghani et Delavar (2014), Sehra et al. (2014), Hayakawa et al. (2012) et Mondzech et Sester (2011). Les contributions de ces travaux consistent principalement en la définition ou l'évaluation de métriques qualité très spécifiques à des problématiques d'étude de données géographiques.

12. `https://www.openstreetmap.org`

Zaveri et al. (2013) ont mené une étude de la qualité de données issues de DBpedia [13] via l'utilisation d'une méthode fondée sur le *crowdsourcing* (voir la description de cette méthode en section 2.1.1) impliquant 58 utilisateurs. Les résultats de cette étude ont permis d'exhiber 17 types de problèmes de qualité des données de DBpedia. Les problèmes les plus fréquents parmi ceux étudiés pour ce jeu de données concernent les dimensions d'*exactitude* et d'*interconnexion*.

Auer et Lehmann (2007) proposent une méthode permettant l'extraction de données du jeu de données Wikipedia [14]. Les données sont extraites puis représentées au format RDF. Quelques problèmes de qualité des données sont décelés à cette occasion, notamment des problèmes d'*exactitude syntaxique* et de *redondance d'information*. Dans ces travaux, les auteurs laissent ouvert le problème d'une évaluation méthodique de la qualité des données.

Böhm et al. (2012) proposent un outil permettant de récupérer et intégrer des données gouvernementales. Un retour d'expérience associé à ce travail pointe les difficultés rencontrées lors de l'exploitation des données. Il est par exemple difficile de choisir les jeux de données pertinents pour l'étude à mener. De notre point de vue, cette difficulté relève non seulement d'un problème de découverte de jeux de données mais également d'un problème de gestion de qualité des données puisqu'il s'agit de savoir associer un niveau de qualité (par exemple de *pertinence*) aux données afin de pouvoir les discriminer. Les auteurs pointent également un manque de documentation associée à certaines données, l'association de documentation aux données pouvant relever des dimensions d'*utilisabilité* et de *compréhensibilité* des données. Les auteurs notent également une grande variété des schémas sous-jacents des données, pointant par là même un niveau de *compréhensibilité* insuffisant de certaines données. Enfin, le niveau de *versatilité* des données pose des problèmes car nécessite de re-paramétrer l'outil régulièrement afin de l'adapter aux données ayant évolué.

Hogan et al. (2010) et Hogan et al. (2012) ont mené des expérimentations leur ayant permis de mettent en évidence divers problèmes de qualité des données dus à des erreurs faites au moment de leur publication. Il est intéressant de noter que les auteurs indiquent à cette occasion que la publication des données est une tâche compliquée, les erreurs n'étant d'ailleurs pas nécessairement dues à une inexpérience des producteurs de données car des producteurs de données très expérimentés (incluant les auteurs eux-mêmes) commettent des erreurs au moment de la publication de leurs données.

3.2 Publication de jeux de données guidée par la qualité

À notre connaissance, il n'existe pas de contribution scientifique issue de la communauté de recherche présentant de cas d'étude d'envergure concernant la gestion de la qualité des données ouvertes du point de vue du producteur de données, utilisant les méta-modèles et frameworks présentés dans la section 2.1.3. Dirschl et al. (2014) sont les seuls auteurs proposant une contribution partielle dans ce cadre. Leurs travaux présentent un cas d'étude d'ouverture de données d'une société d'édition professionnelle, utilisant l'outil *LOD2* (voir section 2.3). Les différentes étapes de la publication y sont présentées. La gestion de la qualité des données y est évoquée, en insistant sur l'importance de la mise en qualité des données, mais les dimen-

13. `http://wiki.dbpedia.org`
14. `http://wikipedia.org`

sions et métriques considérées ne sont pas décrites.

La communauté industrielle est en revanche très active sur cette problématique, comme peuvent le montrer les nombreux groupes de travail actuellement menés par des regroupements d'entreprises. On peut par exemple citer les groupes de travail *Ouverture des données*, *Données personnelles*, *Valorisation des données* et *Architecture et qualité des données* menés dans le cadre de l'association ExQI (2016). Ces groupes de travail ont mené des réflexions sur la gestion de la qualité des données ouvertes, guidées par les enjeux essentiellement stratégiques de l'ouverture des données pour l'entreprise.

De bonnes pratiques pour la publication de données, complémentaires à celles de la littérature, ont été exhibées par Barthélémy et al. (2015). En premier lieu, des réflexions préalables non triviales doivent être menées afin de choisir les données à publier, les méta-données à y associer et le niveau de qualité souhaité. Il convient de définir quelles finalités, quels piliers de valeur ou de l'éthique de l'entreprise seront renforcés par l'ouverture, par exemple, la transparence de l'activité, la responsabilité sociétale, la conformité à la loi, la contribution à l'innovation, le bénéfice économique ou encore l'affichage de performance.

L'ouverture des données étant un investissement, il est également nécessaire de définir les moyens humains, financiers et informatiques qui sont à mettre en œuvre afin de mener cette ouverture. Concernant le coût de la mise en qualité des données, celui-ci peut être discuté. Améliorer la qualité des données a évidemment un coût, mais les données ouvertes sont souvent des données avant tout opérationnelles, utiles et utilisées au sein de l'entreprise. Ainsi, leur mise en qualité est également un bénéfice dans le cadre de leur utilisation en interne à l'entreprise. Dans une vision de la gestion des données au niveau de l'entreprise, il convient donc de ne pas dissocier les deux types d'usages, interne et externe, des données. Dans cet esprit, ouvrir des données, même sous l'impulsion d'une obligation légale au départ, est une opportunité de pouvoir les considérer sous un autre angle et d'améliorer leur qualité dans une double finalité, interne et externe.

Dans le cadre d'une ouverture, les risques liés à l'ouverture (levée de confidentialité par recoupement avec d'autres données, perte d'avantage concurrentiel, dégradation d'image) doivent être évalués au mieux. Puisqu'il est impossible de maîtriser l'usage fait des données, les risques ne peuvent être évalués qu'en cherchant à anticiper les utilisations qui pourraient être faites des données. Cette étape est essentielle mais extrêmement difficile à mener. Il ne pourra malheureusement pas être dressé de liste exhaustive des usages futurs faits des données. Si certains usages "conventionnels" liés au métier couvert peut-être identifiés, d'autres usages enrichissants resteront toujours imprévisibles, par exemple ceux portés par des startups à l'imagination débordante.

Le problème organisationnel, de gouvernance des données en interne de l'entreprise, se pose également de façon plus complexe, l'ouverture des données étant un nouveau volet. La gouvernance des données doit permettre d'évaluer le patrimoine de l'entreprise, d'identifier les données à fort potentiel ou à risques et d'élaborer une stratégie d'ouverture et de mise en qualité répondant aux contraintes réglementaires et aux enjeux de l'entreprise. Il s'agit d'un vaste programme que doivent mener bien les CDO (*Chief data Officer*) dont se dotent de plus en plus d'entreprises.

Les problèmes auxquels les entreprises ou institutions sont confrontées lorsqu'elles décident d'ouvrir une partie de leurs données, incluant la gestion de la qualité de ces données,

sont donc très larges.

Un signe évocateur du manque criant de supports méthodologiques et de retours d'expérience liés à la publication de données est la mise en place de *laboratoires expérimentaux* de publication des données déployés de façon interne dans certaines entreprises. Une bonne pratique consiste à mettre en place une ouverture progressive des données, d'abord à destination d'utilisateurs internes à l'entreprise puis de partenaires extérieurs, en préparation d'une ouverture publique. Cette bonne pratique a été proposée par Barthélémy et al. (2015). Ce processus vise à (i) mieux comprendre les enjeux, pratiques et risques liés à l'ouverture des données tout en limitant les risques associés, et à (ii) améliorer la qualité des données exposées grâce aux retours d'expérience des utilisateurs du laboratoire expérimental.

4 Problèmes ouverts et perspectives

Bien que de nombreux travaux aient été menés, des verrous technologiques et méthodologiques doivent encore être levés. Nous en identifions quelques-uns s'inscrivant dans la mouvance de travaux en cours ou ouvrant la voie à de nouvelles perspectives de recherche.

Méthodologies pour évaluer la qualité et guider l'interprétation des résultats. La définition de la qualité des données dépend de l'utilisation faite de celles-ci. Il est malheureusement très complexe pour un producteur de données ouvertes d'identifier tous les cas possibles d'exploitation des données publiées. Il convient donc aujourd'hui de définir de nouvelles méthodologies de définition de la qualité des données à publier. D'autre part, il peut s'avérer compliqué pour un utilisateur (par exemple un scientifique ou un journaliste des données) d'interpréter les résultats d'une étude qualité car savoir interpréter ces résultats nécessite à la fois une expertise qualité et une expertise métier. Un utilisateur devrait pouvoir être aidé par l'apport d'une expertise humaine ou semi-automatisée. Une piste pourrait être la définition de méthodes permettant la suggestion d'actions de réparation à mener étant donnés certains problèmes qualité exhibés par les résultats d'une étude.

Meta-modèle qualité. D'autres métriques qualité et méthodes d'évaluation associées peuvent être imaginées afin de qualifier la qualité des données ouvertes. Parmi les nombreuses pistes de recherche existantes, celles liées à la corroboration de sources permettant une meilleure évaluation du critère de confiance des sources et celle de l'évaluation de la qualité des données à l'aide de techniques de *crowdsourcing* paraissent particulièrement novatrices et prometteuses. Cette problématique d'évaluation collaborative fait partie intégrante du sujet de la définition d'un meta-modèle général de la qualité des données ouvertes liées.

Dans le cadre de meta-modèle qualité, les modèles de coût liés à la gestion de la qualité des données dans un contexte non ouvert, comme ceux proposés par English (1999) et Eppler et Helfert (2004), doivent être étendus au contexte des données ouvertes.

Interdépendances des dimensions qualité. Comme dans le cadre classique pour lequel le problème reste encore largement ouvert, les interdépendances de dimensions qualités devraient être étudiées. Ce problème est d'ailleurs évoqué par Guéret et al. (2012) qui indiquent de l'enrichissement automatique de liens entre les données peut amener à générer des liens erronés.

En d'autres termes, comme cela a régulièrement été constaté dans le cadre classique, chercher à améliorer une dimension qualité (ici la *complétude*) peut amener à dégrader une autre dimension qualité (ici l'*exactitude*). Les influences négatives entre dimensions sont difficiles à anticiper, des moyens de les gérer devraient être étudiés.

Dans ce contexte, l'interdépendance entre les dimensions qui caractérisent le processus de production de données et celles qui caractérisent la qualité des données liées est particulièrement intéressante. Ces aspects ont été abordés pour des systèmes d'information multisources dans le cadre du projet ANR QUADRIS (Akoka et al. (2007); Berti-Equille et al. (2011)), dont l'un des objectifs était l'étude des interdépendances entre qualité des modèles, des processus et des données. En terme de gestion de la qualité des données, cela aurait un grand intérêt car l'amélioration de la qualité des données peut passer par l'amélioration de leur processus de production.

Cartographier et recommander des jeux de données. Étant donné un besoin, il est très difficile pour un consommateur de données de trouver (puis choisir) les jeux de données ouverts à utiliser. Il est encore nécessaire de pouvoir identifier et cartographier de la façon la plus exhaustive possible les jeux de données disponibles, et de pouvoir recommander des données à un utilisateur, en fonction de son besoin fonctionnel en s'appuyant sur une évaluation de la qualité des données. Un objectif technique visé pourrait être le développement d'une plateforme de grande envergure, à l'image de celles existant actuellement dans le domaine du tourisme. Idéalement, un tel outil devrait permettre de 1) identifier et cartographier les jeux de données disponibles, 2) évaluer la qualité des données de façon collaborative en impliquant des communautés d'intérêt d'utilisateurs et 3) recommander des données à un utilisateur, en fonction de son besoin fonctionnel.

Langages d'interrogation. L'interrogation de données via l'utilisation d'un langage d'interrogation reste d'actualité dans le cadre des données ouvertes liées (par exemple via le langage SPARQL). Dans l'objectif d'une amélioration de l'utilisabilité des données, il serait pertinent de repenser dans ce cadre les travaux préalablement menés pour l'interrogation de données (non liées) guidée par la qualité des données. Au moins deux types de travaux peuvent être considérés : les travaux concernant la définition et la prise en compte de préférences utilisateurs (en particulier les préférences qualité) et les travaux concernant le rapatriement de données multi-sources fondés sur un principe de négociation de contrats qualité (tels que proposés par Berti-Équille (2007)). Un premier pas dans cette direction est le travail mené par Hartig (2009) qui propose une extension du langage SPARQL permettant d'interroger des données RDF porteuses d'information de confiance.

Choix des données et méta-données à publier. Le problème du choix des données et méta-données à publier est spécifique au rôle de producteur de données. La question est ici de déterminer quelles données exposer, en alignement avec les stratégies métier du producteur de données, en particulier si celui-ci évolue dans un environnement fortement concurrentiel ou s'il manipule des données sensibles ou confidentielles (des zones de secret/confidentialité ou d'avantage compétitif pouvant être éventuellement levées par recoupement des données avec d'autres sources). Un autre problème concerne le choix des méta-données à associer aux données, de façon à permettre leur exploitation. Les méthodologies existantes en support au pro-

cessus de publication des données restent relativement rudimentaires et n'apportent pas encore de réponse méthodologique satisfaisante à ces problèmes. Elles doivent être enrichies afin de mieux guider le producteur de données.

Évolution des données. D'autres perspectives de recherche concernent la maintenance des données publiées, ce verrou ayant été identifié dans Zaveri et al. (2014). Quelques problèmes restent ouverts, telles que la maintenance de données pour lesquelles on ne dispose pas d'information de provenance. Un autre problème concerne la gestion de données réparées : comment les stocker, comment prévenir les utilisateurs et sources liées de l'évolution ? Même si des solutions de versionning des données elles-mêmes existent (voir Knuth et al. (2014)), le problème du suivi et de la gestion de la qualité des données au sein de ces outils reste entier.

Certification de données. La grande hétérogénéité de qualité des données ouvertes pose le problème de la certification de leur qualité. Comme souligné lors du DEP ExQI (2015), il est nécessaire de définir des méthodes et organismes de certification des données ouvertes, allant au delà de la simple certification de respect de règle syntaxiques génériques.

5 Conclusion

La publication de données ouvertes éventuellement liées est un phénomène en pleine croissance. Cette immense ressource offre de grandes possibilités d'exploitation. Le niveau très disparate de qualité des données publiées rend cependant leur utilisation difficile, voir risquée. Dans ce cadre, l'évaluation et la maîtrise de la qualité de ces données devient un enjeu de premier plan.

Notre article aborde cette problématique et dresse un état de l'art des approches méthodologiques et techniques de gestion de la qualité des données ouvertes liées proposées dans la littérature. Le périmètre couvert inclut les dimensions et métriques, les frameworks de gestion, les plateformes et outils associés, et enfin les cas d'étude de publication et d'utilisation des données ouvertes liées centrés sur la qualité de celles-ci.

On constate tout d'abord que la plupart des concepts, méthodes et techniques définies dans le cadre classique des données non ouvertes non liées sont au mieux à ré-adapter, lorsqu'ils ne sont pas devenus inadéquats. Un grand nombre de travaux ont donc été proposés pour permettre la gestion des données présentant un caractère lié ou ouvert.

En ce qui concerne l'application des concepts, méthodes et outils existants, on peut noter l'existence de quelques cas d'étude et retours d'expérience « réels » d'utilisation et de publication des données centrés sur la gestion de la qualité des données. Les cas d'étude restent peu nombreux mais montrent clairement le besoin de gestion de la qualité des données ouvertes liées.

Notre étude nous a amenées à recenser des verrous restant encore à lever, relevant de la gestion de la qualité des données du point de vue du producteur de données ou de celui du consommateur de données, les deux points de vue reposant sur des enjeux différents. Les problèmes ouverts sont nombreux, de considérables travaux restent à entreprendre pour permettre d'exploiter toute la richesse des données ouvertes.

Remerciements

Ces travaux ont été partiellement financés par la DGE (Direction Générale des Entreprises) dans le cadre du projet ODIN (Open Data INtelligence), le défi CNRS Mastodons GioQoso et le défi scientifique émergent Université Rennes 1 Quality@PANAM. Les auteurs remercient également l'association EXQI, et en particulier les membres de l'association participant aux groupes de travail et ayant restitué leur vision au cours de la conférence DEP ExQI Paris 2015. Leur point de vue a contribué à l'orientation des travaux présentés ici. Enfin, les auteurs remercient les relecteurs anonymes dont les remarques ont permis d'améliorer la qualité l'article.

Références

Acosta, M., A. Zaveri, E. Simperl, D. Kontokostas, S. Auer, et J. Lehmann (2013). Crowdsourcing linked data quality assessment. In *Proceedings of the International Semantic Web Conference (ISWC)*, pp. 260–276.

Aggrego (consulté en 2016). `http://www.semsoft-corp.com/en/#aggrego-solution`.

Akoka, J., L. Berti-Equille, O. Boucelma, M. Bouzeghoub, I. Comyn-Wattiau, M. Cosquer, V. Goasdoué-Thion, Z. Kedad, S. Nugier, V. Peralta, et S. S. Cherfi (2007). A framework for quality evaluation in data integration systems. In *Proceedings of the International Conference on Enterprise Information Systems (ICEIS)*, pp. 170–175.

Arenas, M., G. I. Diaz, A. Fokoue, A. Kementsietsidis, et K. Srinivas (2014). A principled approach to bridging the gap between graph data and their schemas. *PVLDB 7*(8), 601–612.

Arsanjani, J. J., P. Mooney, A. Zipf, et A. Schauss (2015). Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets. In *OpenStreetMap in GIScience - Experiences, Research, and Applications*, Lecture Notes in Geoinformation and Cartography, pp. 37–58. Springer.

Auer, S., L. Bühmann, C. Dirschl, O. Erling, M. Hausenblas, R. Isele, J. Lehmann, M. Martin, P. N. Mendes, B. V. Nuffelen, C. Stadler, S. Tramp, et H. Williams (2012). Managing the life-cycle of linked data with the LOD2 stack. In *Proceedings of the International Semantic Web Conference (ISWC)*, pp. 1–16.

Auer, S. et J. Lehmann (2007). What have Innsbruck and Leipzig in common ? extracting semantics from wiki content. In *Proceedings of the European Semantic Web Conference (ESWC)*, pp. 503–517.

Barthélémy, N., E. Rossin, G. Aldebert, D. Barrau, G. Rougier, et F. Granovsky (2015). Ouverture des données à l'externe : apprentissage. Compte-rendu du groupe de travail ExQI Open Data.

Basili, V., G. Caldiera, et H. Rombach (1994). The Goal Question Metric Approach. In *Encyclopedia of Software Engineering*. Wiley.

Batini, C., C. Cappiello, C. Francalanci, et A. Maurino (2009). Methodologies for data quality assessment and improvement. *ACM Computing Surveys (CSUR) 41*(3), 16 :1–16 :52.

Batini, C. et M. Scannapieco (2006). *Data Quality : Concepts, Methodologies and Techniques*. Data-Centric Systems and Applications. Springer.

Behkamal, B., M. Kahani, E. Bagheri, et Z. Jeremic (2014). A metrics-driven approach for quality assessment of linked open data. *J. Theor. Appl. Electron. Commer. Res. 9*(2), 64–79.

Berti-Équille, L. (2007). *Information Quality Management : Theory and Practice*, Chapter Quality-Extended Query Processing for Mediation Systems, pp. 23–50. Latif Al-Hakim, IDEA Group Inc.

Berti-Equille, L., I. Comyn-Wattiau, M. Cosquer, Z. Kedad, S. Nugier, V. Peralta, S. S. Cherfi, et V. Thion-Goasdoué (2011). Assessment and analysis of information quality : a multidimensional model and case studies. *IJIQ 2*(4), 300–323.

Bizer, C. et R. Cyganiak (2009). Quality-driven information filtering using the WIQA policy framework. *Web Semantics : Science, Services and Agents on the World Wide Web 7*(1), 1 – 10. The Semantic Web and Policy.

Böhm, C., M. Freitag, A. Heise, C. Lehmann, A. Mascher, F. Naumann, V. Ercegovac, M. Hernandez, P. Haase, et M. Schmidt (2012). GovWILD : Integrating open government data for transparency. In *Proceedings of the International Conference on World Wide Web (WWW)*, pp. 321–324.

Callimachus (consulté en 2016). `http://callimachusproject.org`.

Carroll, J. J., C. Bizer, P. Hayes, et P. Stickler (2005). Named graphs, provenance and trust. In *Proceedings of the International Conference on World Wide Web (WWW)*, pp. 613–622.

Chignard, S. et L.-D. Benyayer (2015). *Datanomics – Les nouveaux business models des données*. FYP Éditions.

CKAN (consulté en 2016). `http://ckan.org/`.

De Feo, J., W. Barnard, et Juran Institute (2005). *Institute's Six Sigma Breakthrough and Beyond - Quality Performance Breakthrough Methods*. McGraw-Hill Professional.

Debattista, J., S. Auer, et C. Lange (2016). Luzzu - A framework for linked data quality assessment. In *Proceedings of the International Conference on Semantic Computing*, pp. 124–131.

Debattista, J., C. Lange, et S. Auer (2014). daQ, an ontology for dataset quality information. In *Proceedings of the Workshop on Linked Data on the Web co-located with the International World Wide Web Conference (WWW)*.

Debattista, J., S. Londoño, C. Lange, et S. Auer (2015). Quality assessment of linked datasets using probabilistic approximation. In *Proceedings of the European Semantic Web Conference (ESWC)*, pp. 221–236.

DEP ExQI (2015). Data Excellence Paris ExQI. Conférence annuelle de l'association ExQI, `http://www.exqi-dep.com`.

Dimou, A., D. Kontokostas, M. Freudenberg, R. Verborgh, J. Lehmann, E. Mannens, S. Hellmann, et R. V. de Walle (2015). Assessing and refining mappings to RDF to improve dataset quality. In *Proceedings of the International Semantic Web Conference (ISWC)*, pp. 133–149.

Dirschl, C., T. Pellegrini, H. Nagy, K. Eck, B. V. Nuffelen, et I. Ermilov (2014). LOD2 for media and publishing. In *Linked Open Data - Creating Knowledge Out of Interlinked Data*

- *Results of the LOD2 Project*, pp. 133–154.

English, L. P. (1999). *Improving Data Warehouse and Business Information Quality*. Wiley and Sons.

Eppler, M. J. et M. Helfert (2004). Classification and analysis of data quality costs. In *Proceedings of the International Conference on Information Quality (IQ)*.

ExQI (Consulté en 2016). Groupes de travail de l'association ExQI. `http://exqi.asso.fr/qui-sommes-nous/objectifs/groupes-dinterets-et-detudes`.

Fan, H., A. Zipf, Q. Fu, et P. Neis (2014). Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science 28*(4), 700–719.

Forghani, M. et M. R. Delavar (2014). A quality study of the OpenStreetMap dataset for Tehran. *ISPRS Int. J. Geo-Information 3*(2), 750–763.

Fürber, C. et M. Hepp (2011a). SWIQA - a semantic web information quality assessment framework. In *Proceedings of the European Conference on Information Systems (ECIS)*.

Fürber, C. et M. Hepp (2011b). Towards a vocabulary for data quality management in semantic web architectures. In *Proceedings of the EDBT/ICDT Workshop on Linked Web Data Management*, pp. 1–8.

Galland, A., S. Abiteboul, A. Marian, et P. Senellart (2010). Corroborating information from disagreeing views. In *Proceedings of the ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 131–140.

GDEF (Consulté en nov. 2015). Global Data Excellence Framework. `http://www.globaldataexcellence.com/en/solutions/data-excellence-framework`.

Gil, Y. et V. Ratnakar (2002). Trusting information sources one citizen at a time. In *Proceedings of the International Semantic Web Conference (ISWC)*, pp. 162–176.

Guéret, C., P. T. Groth, C. Stadler, et J. Lehmann (2012). Assessing linked data mappings using network measures. In *Proceedings of the European Semantic Web Conference (ESWC)*, pp. 87–102.

Hartig, O. (2009). Querying Trust in RDF Data with tSPARQL. In *Proceedings of the European Semantic Web Conference (ESWC)*, pp. 5–20.

Hartig, O. et J. Zhao (2009). Using web data provenance for quality assessment. In *Proceedings of the International Workshop on the role of Semantic Web in Provenance Management (SWPM)*.

Hartig, O. et J. Zhao (2010). Publishing and consuming provenance metadata on the web of linked data. In *Proceedings of the International Provenance and Annotation Workshop (IPAW)*, pp. 78–90.

Hayakawa, T., Y. Imi, et T. Ito (2012). Analysis of quality of data in OpenStreetMap. In *Proceedings of the IEEE International Conference on Commerce and Enterprise Computing (CEC)*, pp. 131–134.

Hogan, A., A. Harth, A. Passant, S. Decker, et A. Polleres (2010). Weaving the pedantic web. In *Proceedings of the WWW2010 Workshop on Linked Data on the Web (LDOW)*.

Hogan, A., J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, et S. Decker (2012). An empirical survey of linked data conformance. *Web Semantics : Science, Services and Agents on the World Wide Web 14*, 14 – 44. Special Issue on Dealing with the Messiness of the Web of Data.

Knap, T., J. Michelfeit, et M. Necaský (2012). Linked open data aggregation : Conflict resolution and aggregate quality. In *Proceedings of the IEEE Computer Software and Applications Conference Workshops (COMPSAC)*, pp. 106–111.

Knuth, M., D. Kontokostas, et H. Sack (2014). Linked data quality : Identifying and tackling the key challenges. In *Proceedings of the Workshop on Linked Data Quality co-located with 10th International Conference on Semantic Systems (LDQ@SEMANTiCS)*.

Kontokostas, D., P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, et A. Zaveri (2014). Test-driven evaluation of linked data quality. In *Proceedings of the International World Wide Web Conference (WWW)*, pp. 747–758.

Kontokostas, D., A. Zaveri, S. Auer, et J. Lehmann. TripleCheckMate : A tool for crowdsourcing the quality assessment of linked data. In *Proceedings of the International Conference on Knowledge Engineering and the Semantic Web (KESW)*.

LOD Catalog (Consulté en octobre 2015). Linked Open Data Catalog. `http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/`.

Mazón, J., J. J. Zubcoff, I. Garrigós, R. Espinosa, et R. Rodríguez (2012). Open business intelligence : on the importance of data quality awareness in user-friendly data mining. In *Proceedings of the Joint EDBT/ICDT Workshops*, pp. 144–147.

Mendes, P. N., H. Mühleisen, et C. Bizer (2012). Sieve : linked data quality assessment and fusion. In *Proceedings of the Joint EDBT/ICDT Workshops*, pp. 116–123.

Mondzech, J. et M. Sester (2011). Quality analysis of OpenStreetMap data based on application needs. *Cartographica 46*(2), 115–125.

OpenDataSoft (consulté en 2016). `http://www.opendatasoft.com/`.

Paulheim, H. et C. Bizer (2014). Improving the quality of linked data using statistical distributions. *Int. J. Semantic Web Inf. Syst. 10*(2), 63–86.

RDF Alerts. RDF Alerts. `https://www.w3.org/2001/sw/wiki/RDF_Alerts`.

RDF Validator. RDF Validator. `http://www.w3.org/RDF/Validator`.

RDFa Developer. RDFa Developer. `https://addons.mozilla.org/en-US/firefox/addon/rdfa-developer/`.

Ruckhaus, E., M. Vidal, S. Castillo, O. Burguillos, et O. Baldizan (2014). Analyzing linked data quality with liquate. In *The Semantic Web : ESWC 2014 Satellite Events*, pp. 488–493.

Schultz, A., A. Matteini, R. Isele, C. Bizer, et C. Becker (2011). LDIF - linked data integration framework. In *Proceedings of the International Workshop on Consuming Linked Data (COLD)*.

Sehra, S. S., J. Singh, et H. S. Rai (2014). A systematic study of OpenStreetMap data quality assessment. In *Proceedings of the International Conference on Information Technology : New Generations, ITNG*, pp. 377–381.

Strong, D. M., Y. W. Lee, et R. Y. Wang (1997). Data quality in context. *Commun. ACM 40*(5), 103–110.

W3C. Converters to rdf. http://www.w3.org/wiki/ConverterToRdf.

W3C. Semantic web related validators. https://www.w3.org/2001/sw/wiki/SWValidators.

W3C Government Linked Data Working Group (2011). The joy of data - cookbook for publishing linked government data on the web. http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook.

W3C Working Group (2014). Best practices for publishing linked data. http://www.w3.org/TR/ld-bp/.

Wang, R. Y. et D. M. Strong (1996). Beyond accuracy : What data quality means to data consumers. *J. Manage. Inf. Syst. 12*(4), 5–33.

Zaveri, A., D. Kontokostas, M. A. Sherif, L. Bühmann, M. Morsey, S. Auer, et J. Lehmann (2013). User-driven quality evaluation of DBpedia. In *Proceedings of the International Conference on Semantic Systems (I-SEMANTICS)*, pp. 97–104.

Zaveri, A., A. Maurino, et L. Berti-Equille (2014). Web data quality : Current state and new challenges. *International Journal on Semantic Web and Information Systems 10*(2), 1–6.

Zaveri, A., A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, et S. Auer (2016). Quality assessment for linked data : A survey. *Semantic Web 7*(1), 63–93. http://www.semantic-web-journal.net/system/files/swj773.pdf.

Summary

The publication of linked data has become a growing phenomenon. This can be explained by the development of web semantic technologies and by regulatory constraints that require from companies and institutions the publication of some of their data of interest. The resulting linked open data sources offer huge opportunities for the development of novel applications. But at the same time, the quality of the data provided by these sources can be poor, making their exploitation and usage difficult, sometimes even risky. Quality issues have to be dealt with in this context. In this paper, we present a state of the art of the approaches proposed in the literature for the management of linked open data quality. The scope of our study covers the dimensions and the metrics, the frameworks, the softwares, and the use cases for quality management of linked open data proposed in literature. Based on this study, we identify some open research problems and perspectives.