La réutilisation des données publiques : quels dispositifs ?

Joumana Boustany\*

Gabriella Salzano\*\*

\* Université Paris Descartes – IUT

143, avenue de Versailles - 75016 Paris

[joumana.boustany@parisdescartes.fr](mailto:joumana.boustany@parisdescartes.fr)

\*\* Université Paris Est Marne-la-Vallée

5, Bd. Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2

[gabriella.salzano@univ-mlv.fr](mailto:gabriella.salzano@univ-mlv.fr)

Membres de l’équipe de recherche DICEN

(Dispositifs d’information et de communication à l’ère numérique)

EA 4420, CNAM

**Résumé.** Les enjeux majeurs des données publiques sont liés à la démocratie, en permettant plus de transparence, et à l’économie, en encourageant l’innovation. Une des conditions facilitant l'exploitation de ces données réside dans la possibilité de les réutiliser. Pour ce faire, il est nécessaire qu’elles soient dans des formats lisibles par la machine et interopérables. Cet article analyse quatre plateformes gouvernementales : celles du Royaume-Uni, des États-Unis, de la Nouvelle-Zélande et de l’Australie. Après avoir étudié des éléments globaux (catégories thématiques, modes d’exploration), une analyse quantitative des différents formats des jeux de données a été réalisée afin de dégager ceux qui sont le plus employés. Ces informations technologiques ont été corrélées à des orientations stratégiques et organisationnelles qui peuvent être complémentaires. L'examen des métadonnées décrivant ces données montre un frein supplémentaire à l’interopérabilité. Ces résultats conduisent à préconiser des actions facilitant la montée en charge et le développement de ces plateformes.

**Abstract.** The main stakes of public data are linked with democracy, to bring more transparency, and with economy, to encourage innovation. One of the conditions which makes the exploitation of this data easier is its reusability. To do so, it is necessary that these datas are in formats readable by the machine and are also interoperable. This article analyzes four platforms developed at government level, such as in the United Kingdom, United States, Australia and New Zealand. After having studied their global elements (thematic categories, exploration modes, ...), a quantitative analysis of the data sets’ different formats has been realized to identify the most used. These elements have been correlated with some organizational and strategic choices, which may be complementary. A brief analysis of studied platforms metadata shows supplementary obstacles to interoperability. These results lead us to recommend actions able to make the scalability and development of these platforms easier.

1. Introduction

La réutilisation des données publiques est une problématique récurrente dans différents pays du monde. En effet, la libération des informations du secteur public (ISP) présente des enjeux sociaux, culturels et économiques. Ces ISP constituent une source d’information à valeur ajoutée pouvant être exploitées par les entreprises et contribuant ainsi à la croissance économique et à la création d’emplois.

Les instances dirigeantes sont conscientes de ces opportunités et tentent de réguler ce marché. C’est le cas de l’Union européenne (UE) avec la directive 2003/98/CE du 17 novembre 2003, concernant la réutilisation des informations du secteur public. Elle vise à faciliter la réutilisation transfrontalière des données en fixant « *un ensemble minimal de règles* » sans aucune contrainte d’application. Il en est de même de la directive 2003/4/CE relative à l’accès du public à l’information en matière d’environnement ainsi que sa diffusion et de la Directive INSPIRE, entrée en vigueur le 15 mai 2007.

Aux États-Unis, le Président Obama a exigé que toutes les administrations américaines anticipent sur les réclamations des citoyens par une diffusion proactive des informations en utilisant les nouvelles technologies. Depuis, la plateforme *Data.gov* a vu le jour rendant ainsi accessible une série de données du gouvernement américain. D’autres pays ont suivi comme le Royaume-Uni, la Nouvelle-Zélande et l’Australie. La France a initié la création du portail data.gouv.fr le 22 février 2011.

* 1. Objectifs de l’étude

Une analyse détaillée des textes réglementaires a permis de constater que, sans imposer un format de données spécifique, la plupart des États s’accordent sur la nécessité de définir des formats ouverts indépendants de la plateforme, lisibles par la machine et accessibles au public[[1]](#footnote-1). En effet, l’une des barrières à la réutilisation de ces données réside dans « *le manque de standards et de métadonnées communs et dans l’absence de principes communs pour stocker et décrire l’information* »[[2]](#footnote-2). Compte tenu de ce constat, la problématique de cette recherche consiste à apporter des éléments de réponse aux questions suivantes : quels sont les formats les plus utilisés ? Existe-t-il des formats plus adaptés à la diffusion et à la réutilisation des données publiques ? Dans quelle mesure ces dispositifs s'insèrent-ils dans des objectifs d'interopérabilité ?

* 1. Corpus et méthodologie

Les catalogues qui recensent les portails mettant à disposition des données publiques sont assez nombreux. La *Semantic Web Company* utilise la cartographie de Google pour les présenter. Elle les classe en cinq catégories :

1. Autorités gouvernementales locales ou régionales : Paris, Londres, ...
2. Initiatives locales ou régionales d’ordre privé : Berlin, Colorado, …
3. Autorités gouvernementales nationales : data.gov, data.gov.uk, …
4. Initiatives privées nationales : opengovdata.ru, datagov.it, …
5. Initiatives transnationales : Banque mondiale, OCDE, …

Cette étude analyse des plateformes pilotées par des autorités gouvernementales nationales (Type 3). Ces dernières disposent d’un pouvoir institutionnel et peuvent imposer leur choix à l’échelle d’un pays, ce qui n’est pas le cas des autres initiatives. Ainsi ont été sélectionnées quatre plateformes : [data.gov](http://www.google.com/url?q=http%3A%2F%2Fdata.gov&sa=D&sntz=1&usg=AFQjCNFA6mHV3awfDO42uTwaYOenE0SGng) (États-Unis, mai 2009), [data.govt.nz](http://www.google.com/url?q=http%3A%2F%2Fdata.govt.nz&sa=D&sntz=1&usg=AFQjCNHQv6tO8pdVDGzbeQvf4u1SyiTJ-A) (Nouvelle Zélande, novembre 2009), [data.gov.uk](http://www.google.com/url?q=http%3A%2F%2Fdata.gov.uk&sa=D&sntz=1&usg=AFQjCNGlekx3eAQutEU6rVXfddLacqGb0A) (Royaume-Uni, janvier 2010) et [data.gov.au](http://www.google.com/url?q=http%3A%2F%2Fdata.gov.au&sa=D&sntz=1&usg=AFQjCNENv4_vRDVw-qv3glbQmB-LgoKTCQ) (Australie, mars 2011). Elles seront désignées par la suite avec les acronymes (US), (NZ), (UK) et (AU).

La démarche adoptée s’appuie sur l’élaboration d’une grille comparative. Elle se compose de métadonnées conformes au Dublin Core[[3]](#footnote-3) (URL, titre de la plateforme, date de création, etc.), ainsi que de métadonnées pour recueillir des informations qualitatives et quantitatives, notamment concernant les formats. Compte tenu de la disparité entre les approches adoptées par ces portails (volume des données, mode de navigation, etc.), il était nécessaire d’obtenir des données quantitatives servant essentiellement à dégager des tendances et des ordres de grandeur. Ces éléments ont été obtenus par des interrogations ou filtrages pour les plateformes (US) et (UK), et par des calculs manuels pour (NK) et (AU).

1. Résultats[[4]](#footnote-4)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Plateforme** | **US** | **NZ** | **UK** | **AU** |
| Volume | 3259 | 521 | 5600 | 230 |
| BD | - | 18 | 3 | - |
| CAD (DWG) | - | ***80*** | - | - |
| CSV | 2252 | 149 | 1017 | 154 |
| GIS | - | ***116*** | - | - |
| HTML | - | 66 | 56 | ***4*** |
| JSON | - | - | 386 | - |
| KML/KMZ | 22 | 107 | - | 19 |
| PDF | ***9*** | 97 | 78 | 33 |
| RDF | 302 | - | 426 | - |
| SHP | 172 | - | 4 | 6 |
| XLS | ***521*** | 290 | 511 | - |
| XML | 317 | 29 | 418 | 15 |
| ZIP | - | - | 6 | ***10*** |

Les résultats concernent uniquement les données brutes publiées par les plateformes, sans prendre en compte les données calculées, par exemple des cartes accessibles par des applications géographiques, présentes dans (US) et (UK). Sur toutes les plateformes, sauf (UK), la somme des comptages, tous formats confondus, est supérieure au nombre de jeux de données affiché par la plateforme. Ceci s’explique par le fait qu’un même jeu de données est proposé en plusieurs formats. Dans le cas de (UK), cette considération doit probablement se balancer avec deux autres : des jeux de données peuvent en contenir d’autres (non comptabilisés à un premier niveau), présence du format compressé zip. L’analyse des différentes plateformes a permis de recenser les formats présentés dans le tableau ci-dessus, limité aux formats ayant un nombre significatif d’occurrences[[5]](#footnote-5). Les chiffres en gras italiques correspondent à des formats regroupés par les plateformes dans une classe « Autre ».

* 1. **Des formats hétérogènes**

Les formats de données sont très variés. Les plus largement proposés sont le Text/CSV (comprenant XLS dans (AU)), suivi par XLS, XML et RDF. Certains formats ne sont pas adoptés par toutes les plateformes : BD est utilisé seulement dans (UK) et (NZ) ; ZIP seulement dans (UK) et (AU) ; RDF seulement dans (UK) et (US) ; et JSON seulement dans (UK). HTML ne figure pas dans (US) et les formats géographiques CAD et GIS sont spécifiques à (NZ). Plus précisément, les formats géographiques, absents des données brutes de (UK), sont présents dans les applications (UK), mais non explorées dans cette étude.

L’importance que les plateformes accordent à ces formats est également variable à l’exception de (UK) qui les met au même niveau. CSV, XLS, XLM, PDF sont présents partout, mais CSV et XLS sont regroupés dans (AU), XLS (US), PDF (US et AU), CAD et GIS (NZ) sont classées dans une catégorie « Autres ».

La plateforme (UK) a nécessité un traitement supplémentaire. En effet, le filtrage de ses jeux de données sur les formats laisse apparaître des problèmes de classification : par exemple les fichiers Excel sont classés sous XLS, xls, excel et les fichiers CSV et csv sont séparés.

* 1. Des métadonnées hétérogènes

En ce qui concerne les métadonnées[[6]](#footnote-6), la comparaison des jeux de métadonnées sur quatre sets de données de la même catégorie (Health) montre globalement un grand nombre de correspondances : a) la couverture thématique, représentée par les keywords (US), (AU), (NZ) et par les tags (UK), ainsi que la couverture temporelle (création, publication, mise à jour, fréquence) sont toujours présents ; b) la couverture spatiale est présente dans tous les jeux, à l’exception de (NZ) ; c) la licence est mentionnée seulement dans (UK) et (AU) ; d) seulement (US) indique des éléments de Dataset metrics (nombre de téléchargements) ; e) la participation des internautes est sollicitée/attestée par différents champs : *rating* (US, AU), *Give feedback to department* (UK), *Leave a Reply* (AU), *Suggest keyword* (2-NZ).

Cependant, les très nombreuses correspondances s’accompagnent de conflits structurels, voire sémantiques, notamment pour la couverture thématique.

1. Choix pour l’interopérabilité

L’interopérabilité peut être définie comme la capacité de deux ou plusieurs systèmes à échanger des données et partager des informations. L’interopérabilité des formats est préalable à celle requise aux niveaux « supérieurs » : structurel, sémantique, d’organisation de connaissances (Sheth, 1999). Tous les niveaux, y compris celui des formats, comprennent plusieurs dimensions, en plus de celle technologique : chacun comporte des choix stratégiques, industriels et organisationnels (politiques de standardisation, investissements, formations...), opérés par les gouvernements, les industries du logiciel et les entreprises utilisatrices (Tsilas, 2007).

Dans une perspective d’interopérabilité, cette discussion focalise sur les formats proposés par les plateformes étudiées.

Le format CSV est adapté pour une diffusion rapide de données, car il est généralement disponible chez les producteurs et facilement mis en place à partir d’outils bureautiques (Excel).

RDF est un format exploitable automatiquement, lisible par la machine, utilisé pour publier des jeux de données adressables sur le Web par des URIs. En utilisant ces URIs, d’autres données peuvent « pointer » sur celles-ci. L’utilisation de RDF et du langage associé SPARQL permettent de créer un véritable réseau de données, *linked data* (Bizer, Heath & Berners-Lee, 2009, Gandon, 2009), et apporter aux utilisateurs plus d’informations que les sets de données initiaux.

(UK) préconise CSV comme format de base pour publier des données (« *we are where we are* ») et en même temps elle met en valeur les *linked data*, en les regroupant dans un onglet spécifique de même nom[[7]](#footnote-7). Celui-ci liste plusieurs points de départ (*starting points*) pour initier la navigation, et des tutoriaux. Toutefois, la même plateforme souligne que RDF et SPARQL ne peuvent pas être utilisés tels quels par des utilisateurs non-développeurs. Elle met donc à la disposition des utilisateurs une interface de programmation pour interroger les *linked data* et générer des formats divers (XML, JSON, HTML,…). Cette interface, illustrée sur le *starting point* Education, est configurable et adaptable à d’autres points d’entrée et permet de naviguer dans un graphe de *linked data*, créer des vues et des filtres et générer des impressions.

1. Conclusion

Les choix des formats reflètent des orientations stratégiques et organisationnelles, ainsi que technologiques. Souvent, ces choix ne sont pas antinomiques, mais complémentaires, comme le prouve le fait qu’un même jeu de données est diffusé dans plusieurs formats.

Par exemple, le format CSV apparaît opportun pour initier rapidement et à un coût minime une dynamique de réutilisation de données publiques. En termes d’interopérabilité, ce choix a une portée immédiate et très large au niveau des entreprises disposant de jeux de données avec des formats compatibles, issus notamment de plateformes bureautiques. Ce format semble éligible par rapport au critère d’adoption défini par la *Library of Congress*, qui se réfère au degré avec lequel le format est déjà utilisé par les créateurs, les diffuseurs ou les utilisateurs des ressources d’information. (Library of Congress, 2007).

Le format RDF, en reliant des jeux de données, apporte un potentiel de réutilisation des données publiques bien plus élevé. Cependant l’adoption de ce format requiert des compétences spécialisées et pointues au sein des entreprises et relève actuellement du domaine de la recherche. Le choix de ce format apporte une contribution certaine à la communauté du Web sémantique, pour la représentation et intégration d’informations issues de plateformes hétérogènes (Shadbolt et al. 2011).

L’étude sur les formats utilisés par les quatre plateformes a été complétée par une analyse de métadonnées. Celles-ci constituent des éléments essentiels des infrastructures d’interopérabilité, notamment pour les activités de classification et de découverte des données. Les métadonnées, homogènes à l’intérieur de chaque plateforme, sont hétérogènes d’une plateforme à l’autre. Ceci constitue un frein à l’interopérabilité dans des contextes à large échelle (Guijarro, 2007), qui devrait être garantie dès à présent, pour le long terme (Karasti et al. 2010).

Plusieurs initiatives peuvent contribuer à la diffusion et à la montée en puissance de ces plateformes, dans un contexte de plus en plus « globalisé », pour les entreprises comme pour les citoyens : a) solliciter des utilisateurs quant à la réutilisation effective de données publiques, à l’échelon national et international ; b) harmoniser des classifications et métadonnées par les différents pays ; c) concevoir, développer et diffuser des applications d’intégration de données publiques à l’échelle internationale.

1. Glossaire

|  |  |
| --- | --- |
| CAD | Computer-aided design |
| CSV | Comma-separated values |
| DOC | Abréviation de document. Format de fichier propriétaire |
| DWG | DraWinG. Format natif des fichiers générés par le logiciel AUTOCAD |
| GIS | Geographic Information System |
| HTML | Hyper-Text Markup Language |
| JSON | JavaScript Object Notation. Format léger d'échange de données |
| KML | Keyhole Markup Language. Norme du Open Geospatial Consortium |
| KMZ | Les fichiers KMZ sont des fichiers .KML compressés |
| PDF | Portable Document File. Document multi-plateformes créé par Adobe Acrobat |
| RDF | Resource Description Framework. Fichier contenant des informations sur un site web dans un format structuré appelées des métadonnées |
| SHP | Abrégé de *Shapefile.* Un format ouvert issu du monde des Systèmes d'Informations Géographiques |
| SPARQL | Recommandation du W3C. Langage de requête pour les RDF |
| URI | Uniform Resource Identifier |
| XLS | Format de fichier propriétaire pour tableur |
| XML | Extensible Markup Language |
| ZIP | Format de fichier permettant l'archivage et la compression de données sans perte de qualité. |

1. Références

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22.

Gandon, F. L. (Ed.). (2009). IC 2009 : Actes des 20es Journées Francophones d’Ingénierie des Connaissances « Connaissance et communautés en ligne » du 25 au 29 mai 2009 à Hammamet, Tunisie avec la Plate-forme AFIA (xx-312 p.). Grenoble: PUG.

Guijarro, L. (2007). Interoperability frameworks and enterprise architectures in e-government initiatives. in Europe and the United States. Government Information Quarterly, 24(1), pp. 89-101.

Karasti, H., Baker, K. S., Millerand, F. (2010). *Infrastructure Time: Long-Term Matters in Collaborative Development.* Computer Supported Cooperative Work : An International Journal 19: 377-415. Special Issue on Supporting Scientific Collaboration Through Cyberinfrastructure and e-Science.

Library of Congress. (2007). *Sustainability of Digital Formats : Planning for Library of Congress Collections*. <http://www.digitalpreservation.gov/formats/sustain/sustain.shtml#adoption>

Shadbolt, N., Salvadores, M., & Alani, H. (2011). eGovernment. *Handbook of Semantic Web Technologies*. [http://eprints.ecs.soton.ac.uk/21711/](http://www.google.com/url?q=http%3A%2F%2Feprints.ecs.soton.ac.uk%2F21711%2F&sa=D&sntz=1&usg=AFQjCNF6ku9NbvOzBg1TmPj01cJuhD9myQ)

Sheth, A. (1999). Changing Focus on Interoperability in Information Systems: From System, Syntax, Structure to Semantics. *The Kluwer international series in engineering and computer science. VLSI, computer architecture and digital signal processing*, *5*(495), 5–29.

Tsilas, N. L. (2007). Enabling open innovation and interoperability: recommendations for policy-makers. Proceedings of the 1st international conference on Theory and practice of electronic governance (p. 53–56). ACM.

1. [Un format est dit ouvert si son mode de représentation a été rendu public par son auteur et s’il est libre d’utilisation. Il est lisible par la machine (machine readable),](http://www.google.com/url?q=http%3A%2F%2Fwww.opengeospatial.org&sa=D&sntz=1&usg=AFQjCNFVyh1JPyAcXm-xWzHCEXzyJLklcg) s’il peut être interprété automatiquement par des programmes informatiques. [↑](#footnote-ref-1)
2. European commission : Information society. Barriers to the Re-use of PSI. [http://ec.europa.eu/information\_society/policy/psi/barriers/index\_en.htm](http://www.google.com/url?q=http%3A%2F%2Fec.europa.eu%2Finformation_society%2Fpolicy%2Fpsi%2Fbarriers%2Findex_en.htm&sa=D&sntz=1&usg=AFQjCNGhGrrgBT1e0qsvh96StoI6Jy-KWA) [↑](#footnote-ref-2)
3. [Dublin Core Metadata Initiative.](http://www.google.com/url?q=http%3A%2F%2Fwww.opengeospatial.org&sa=D&sntz=1&usg=AFQjCNFVyh1JPyAcXm-xWzHCEXzyJLklcg) [http://dublincore.org](http://www.google.com/url?q=http%3A%2F%2Fdublincore.org%2F&sa=D&sntz=1&usg=AFQjCNHc99lxIzXFMIljRA80fBMutr0soA) [↑](#footnote-ref-3)
4. Pour les sigles, voir glossaire. [↑](#footnote-ref-4)
5. Exemple de résultats non significatifs : (NZ) 1 ASCI, 1 Datastream, 1 DOC ; (AU) 1 TXT ; (UK) 8 DOC, 3 TIFF, 1 GIF, 12 PPT[.](http://www.google.com/url?q=http%3A%2F%2Fwww.opengeospatial.org&sa=D&sntz=1&usg=AFQjCNFVyh1JPyAcXm-xWzHCEXzyJLklcg) [↑](#footnote-ref-5)
6. Les métadonnées des sets de données (UK) sont téléchargeables, aux formats CSV et JSON : [http](file:///C:\Users\Documents%20and%20Settings\SALZANO.PORTMLVIGM01\Local%20Settings\Temp\http)[://data.gov.uk/search/apachesolr\_search/metadata](http://www.google.com/url?q=http%3A%2F%2Fdata.gov.uk%2Fsearch%2Fapachesolr_search%2Fmetadata&sa=D&sntz=1&usg=AFQjCNH9znJwKfR_xCnu_K2-mI-dHG0nlg) [↑](#footnote-ref-6)
7. data.gov.uk - Formats pour Linked data : <http://data.gov.uk/blog/guest-post-developers-guide-linked-data-apis-jeni-tennison> [↑](#footnote-ref-7)