OPEN DATA ET EGOVERNMENT

ISABELLE BOYDENS

Résumé –

Cette note propose un regard critique sur le phénomène des « *open data* » dans l'eGovernment, à l'heure où une directive européenne récemment parue en 2013 en encourage la diffusion. Une question cruciale sera omniprésente : la qualité des « *open data* » eu égard à leurs usages attendus, qualité potentiellement douteuse, dans la mesure où celles-ci prolifèrent dans l'environnement ouvert et non contrôlé du web. Après une définition du concept d'*open data* sur la base d'exemples, l'objet de la directive ainsi que ses implications en termes financiers et juridiques sont évoqués. Une typologie des *open data*, selon que celles-ci sont publiques, statiques ou dynamiques, plus ou moins contrôlées à la source ou restreintes à un public donné est ensuite évaluée. Les aspects fonctionnels des *open data*, depuis les formats jusqu'aux normes d'interopérabilité sémantique et à la mise en place de plateformes de « *citizen engagement* » sont alors évoqués et suivis, en guise de conclusion, de plusieurs recommandations pratiques.

Abstract –

Deze nota biedt een kritische blik op het fenomeen "*open data*" in het eGovernment op het moment dat een Europese richtlijn, die onlangs in 2013 verscheen, de verspreiding ervan aanmoedigt. Een cruciale kwestie zal alomtegenwoordig zijn: de kwaliteit van de "*open data*" hangt af van het verwacht gebruik er van en is eerder onzeker, aangezien "*open data*" vooral gebruikt worden in de open en niet-gecontroleerde omgeving van het web. Na een definitie van het concept *open data* aan de hand van voorbeelden worden het doel van de richtlijn alsook de implicaties ervan op financieel en juridisch vlak besproken. Daarna wordt een typologie van *open data* geëvalueerd naarmate deze een publiek, statisch of dynamisch karakter hebben en naargelang ze in mindere of meerdere mate aan de bron gecontroleerd worden of beperkt worden tot een bepaald publiek. Vervolgens komen de functionele aspecten van *open data* aan bod, gaande van de formaten tot de semantische interoperabiliteitsnormen en de opstelling van "*citizen engagement*"-platformen. Tot slot zullen er meerdere praktische aanbevelingen gegeven worden.

Table des matières

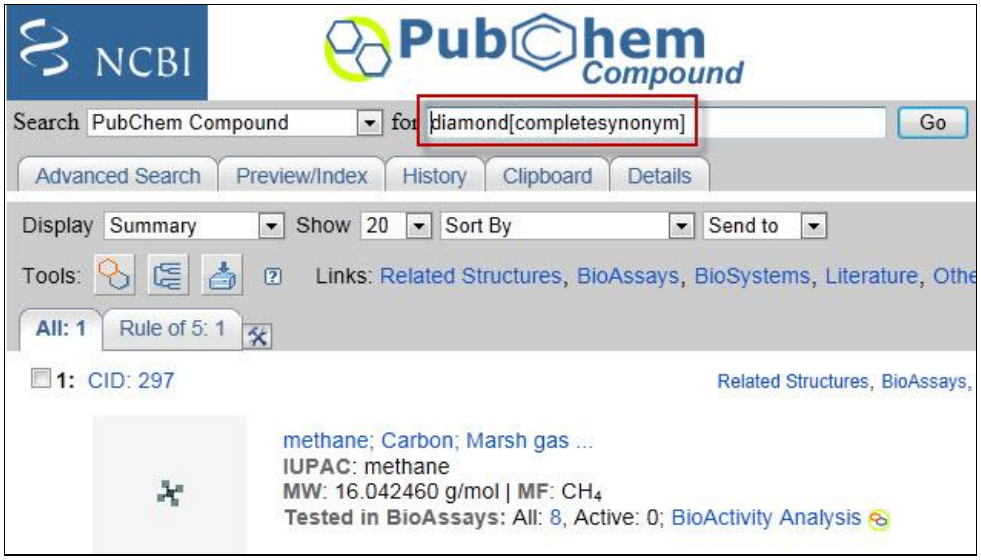
1. Introduction : définition et exemples

L'*open data* est un mouvement né aux États-Unis[1] où il est officialisé depuis la loi « *Freedom of Information Act* » de 1966. Le mouvement a pris de l'ampleur à la fin des années 1990, avec l'émergence d'Internet.

La notion d'*open data* est très large. La Commission européenne en fournit la définition suivante : « *Open data refers to the idea that certain data should be freely available for use and re-use. »*[2]. Nous verrons plus loin que l'on trouve dans la pratique des variantes importantes par rapport à cette définition qui soulève de nombreuses questions juridiques, éthiques, financières, qualitatives et sémantiques, notamment.

Dans ce rapport nous nous intéressons principalement au secteur de l'eGovernment, mais la notion de donnée ouverte concerne aussi le monde scientifique, par exemple, à travers la diffusion des sources à la base des expériences scientifiques. Citons *PubChem,* base de données des molécules chimiques recensant plusieurs dizaines de millions d'entrées diffusées en accès libre (Figure 1). L'alimentation de la base est contrôlée par la *National Library of Medecine* aux USA, ce qui en garantit la fiabilité. L'initiative a toutefois fait l'objet de vives protestations de la part des firmes privées qui diffusaient ces informations préalablement à des fins lucratives.

Figure 1. PubChem : écran de recherche



[1]ASSAR S., BOUGHZALA I. et BOYDENS I., « Back to Practice : a Decade of Research in eGovernment » . In ASSAR S., BOUGHZALA I. et BOYDENS I., éds., « Practical Studies in E-Government : Best Practices from Around the World» , New York, Springer, 2011, p. 1-12 (chapitre 1). Voir aussi, avec les précautions critiques d'usage : http://en.wikipedia.org/wiki/Open_data

[2]http://ec.europa.eu/digital-agenda/en/open-data-0. L'*Open Knowledge Foundation* en fournit une définition similaire « *A piece of data or content is open if anyone is free to use, reuse, and redistribute it* » (http://opendefinition.org/od/).

Née en 2001 dans la communauté des informaticiens et traversée par l'idéal encyclopédiste et universaliste du XIXème siècle, *Wikipédia* constitue un autre exemple bien connu d'information ouverte au grand public et alimentée par celui-ci.

Le projet repose sur plusieurs bonnes pratiques faisant appel au savoir-vivre des contributeurs et à la neutralité de leur point de vue. L'encyclopédie multilingue, dont la qualité est nécessairement inégale, se développe ainsi de manière collaborative, l'historique de chaque modification permettant un contrôle potentiellement rapide des opérations individuelles de vandalisme, par exemple[3].

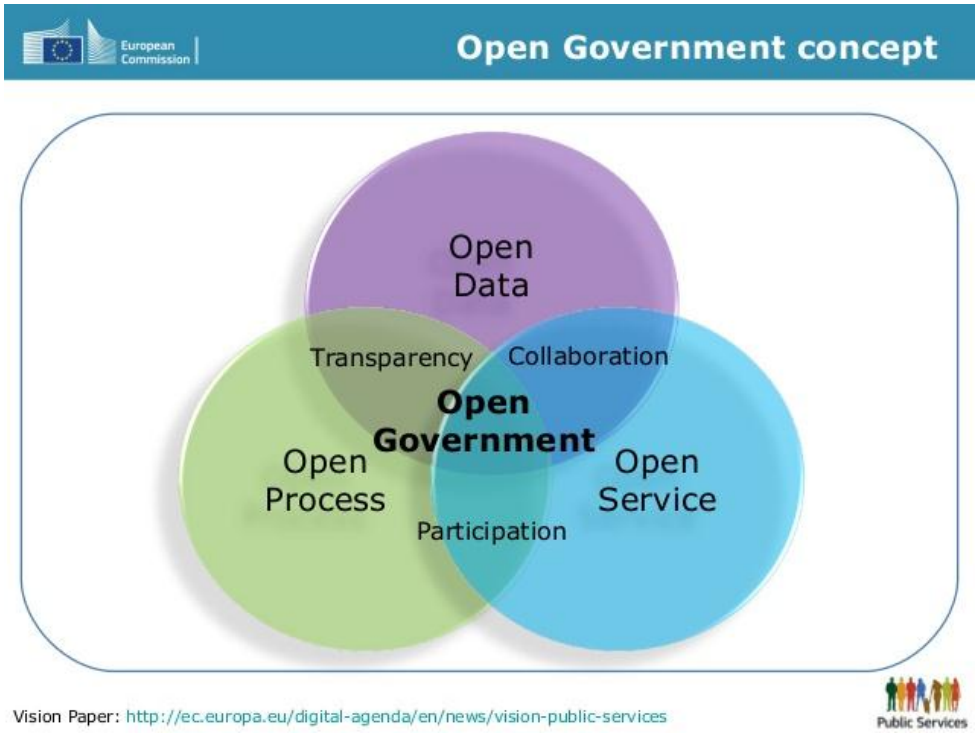
Comme l'illustre la Figure 2, les *open data* constituent un sous-ensemble de l'Open Government, dont l'objet consiste en la mise en ligne de processus et services transversaux au sein de l'administration, mais aussi entre celle-ci et ses partenaires, dont les citoyens. Les *open data* ont pour objectif très général de produire de la valeur en vue :

- de la création de nouveaux services ;
- de favoriser l'innovation (à travers l'émergence d'usages nouveaux a priori inattendus ou inconnus) ;
- d'obtenir des gains via le partage collaboratif de données au sein et à travers les administrations publiques ou toute instance externe ;
- d'encourager la participation des citoyens dans la vie sociale et politique et d'accroître la transparence du gouvernement.

En ouvrant les données, on espère en effet que dans un cadre donné, les citoyens ou toute instance externe en fassent spontanément usage de manière à créer des applications collaboratives porteuses de valeur ajoutée. Ces usages sont a priori imprédictibles, mais nous verrons que plusieurs exemples de réalisations existent déjà.

[3] MOATI A. et BACHELET R., « Wikipedia, un projet hors normes ? » In CARRIEU-COSTA M.-J., BRYDEN A. et COUVEINHES P. éds, Les Annales des Mines, Série « Responsabilité et Environnement » (numéro thématique : « La normalisation : principes, histoire, évolutions et perspectives »), Paris, n° 67, juillet 2012, p. 48-53.

Figure 2. Schéma de l'Open Government (C. E.)



La question des « *open data* » est en 2014 à nouveau d'actualité sur le plan réglementaire international en matière d'eGovernment, mais aussi sur celui des stratégies pratiques actuellement envisagées dans l'informatique administrative. Aussi est-il intéressant de se pencher sur ce phénomène et ses composantes en vue d'en évaluer les différents développements possibles et de formuler un ensemble de recommandations pratiques.

Sur le plan européen, le 26 juin 2013, la directive européenne PSI (*Public Sector Information*) de 2003 encourageant la diffusion des « *open data* », a en effet fait l'objet d'une mise à jour à transposer par les États-membres pour le 18 juillet 2015 (*Directive 2013/37/EU of the European Parliament and of the Council of 26 June 2013*), ci-après directive PSI 2013[4]. Nous en envisageons les grandes lignes au point suivant. Par ailleurs, les Nations Unies, dans le rapport « *Trends from the UN 2014 e-Government Survey* », retiennent dans l'un de ses six thèmes stratégiques « *l'Open Government Data* »[5].

Sur le plan pratique et stratégique de l'IT, Gartner a retenu en 2014 l'Open Government dans son planning de recherche[6], en vue d'en évaluer la

[4] *Journal Officiel de L'Union Européenne*, 27 juin 2013, L 175/1-8. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:175:0001:0008:FR:PDF

[5] KERBY, R., « Trends from the UN 2014 e-Government Survey » , Data days (http://www.datadays.eu/), Ghent, 17-19 février 2014, http://www.slideshare.net/SarahBuelens/kerby-31618205

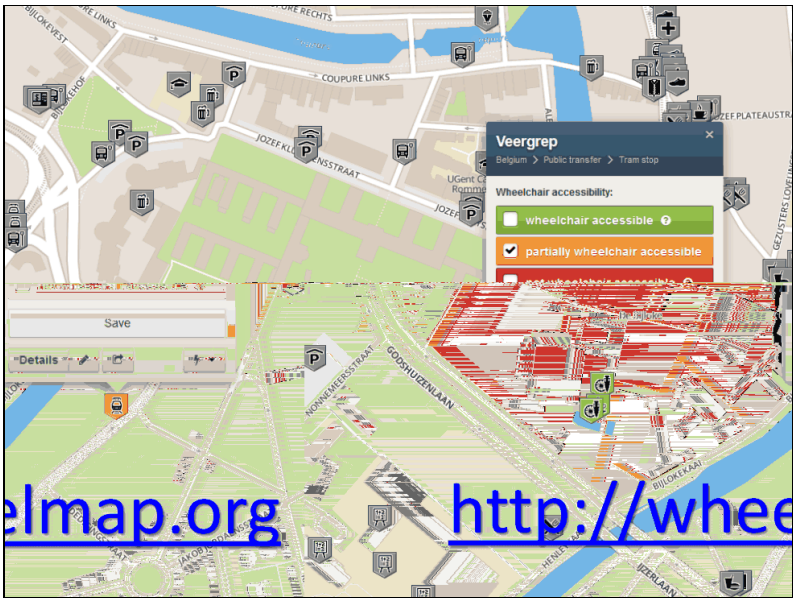
[6] HOWARD R., « Agenda Overview for Government 2014 » , Gartner, 3 janvier 2014, G00257797.

valeur (sur la base de métriques et de benchmarks), et non sans en avoir envisagé les faiblesses, liées au manque de qualité des *open data* publiées sur Internet et dont la mise à jour peut s'avérer très erratique[7].

De nos jours, dans le contexte de la crise, les administrations publiques souhaitent par ailleurs répondre au mouvement du W2.0 en diffusant elles-mêmes des données « ouvertes » dans l'espoir qu'elles fassent l'objet d'applications coopératives libres créatrices de valeur (mashups). Par ailleurs, alors que les entreprises, citoyens, employeurs… ne cessent de communiquer de l'information aux administrations, l'idée est que ces dernières fournissent également en retour, à travers le mouvement « *open data* », des informations à valeur ajoutée (en plus de leur mission première, en tant qu'élément constitutif de l'appareil d'État : prélèvement et redistribution des contributions, exécution de services au profit des administrés, application de la loi, etc.).

Un état de l'art de la question a été récemment présenté à Gand en février 2014 lors des Data Days (www.datadays.eu). La figure 3 présente un exemple d'*open data* très classique, dans le cadre des « *smart cities* », permettant de localiser et d'évaluer dans la ville de Gand les points d'accès pour les personnes handicapées circulant en chaise roulante[8].

Figure 3. The benefits of open public transport data and possible applications



Dans la suite du document, nous proposons d'aborder la notion d'*open data* telle qu'envisagée par la directive européenne PSI de 2013 (point 2), incluant les questions qu'elle soulève et notamment les impacts juridiques

[7] LANEY D., BUYTENDIJK F. et LINDEN A., « Predicts 2014 : Innovating with Information Will Demand New Data, Organizations and Ideas ». Gartner, 29 novembre 2013, ID:G00259174, p. 8.

[8] ABELSHAUSEN B., « *The benefits of open public transport data and possible applications ».* (http://www.datadays.eu/), Ghent, 17-19 février 2014. https://docs.google.com/file/d/0B-uNSfOQ-TR_S1ZzaElsZ1NscGM/edit?pli=1

et en termes de licences ainsi que les impacts en termes de coûts, « open », ne signifiant pas nécessairement « free of charge ». Nous présentons ensuite une typologie des « *open data* » (point 3), selon qu'elles sont statiques (statistiques), de type « *real time* » (trafic…), ciblées sur un public donné, « *open* » ne signifiant pas nécessairement « public » et, enfin, en fonction de leur gestion plus ou moins contrôlée. Les aspects fonctionnels associés à la mise en place concrète des « *open data* », depuis les formats et standards, jusqu'aux plateformes de visualisation et d'échange sont envisagés au point 4. En guise de conclusion, le rapport se termine par un ensemble de recommandations pratiques (point 5).

2. La directive européenne PSI 2013

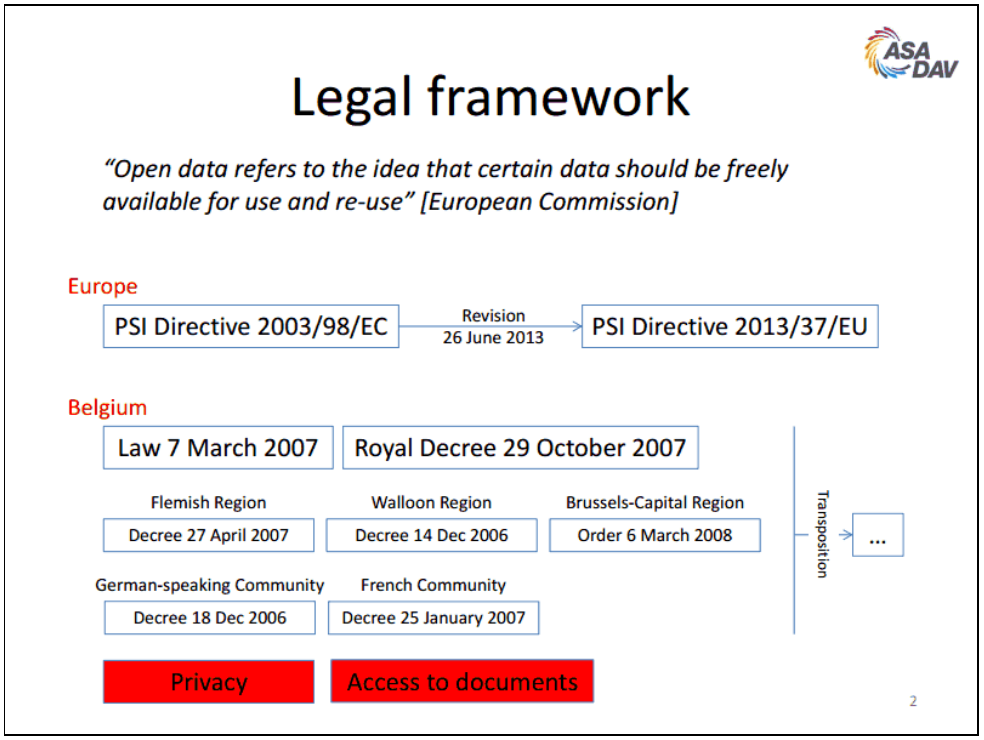
Nous présentons ci-dessous les grandes lignes de la mise à jour de la directive européenne PSI 2013 ainsi que les questions qu'elle soulève sur le plan financier et des licences (respect du droit d'auteur (*copyright)*, protection de la vie privée (*privacy)*, prévention des risques en cas d'exploitation de données de qualité inadéquate…).

2.1. Objet de la directive PSI 2013

Devant être transposée dans le droit interne de chaque État-membre pour le 18 juillet 2015, la directive de 2013[9] constitue une mise à jour de celle de 2003 (Figure 4) qui avait déjà donné le jour à une plateforme de données ouvertes au niveau européen (par exemple : http://open-data.europa.eu/en/data/) et au sein de chaque État-membre, par exemple, pour la Belgique, le site géré par l'ASA-DAV (http://publicdata.belgium.be/fr) ou celui géré par Fedict (http://data.gov.be/), mais aussi de certaines régions ou villes. Nous en présentons ici un aperçu générique qui demandera un approfondissement par des juristes de formation.

[9] À la différence d'un règlement européen qui est d'application directe pour chaque État-membre, une directive nécessite une transposition en droit interne par des institutions législatives et donc, des débats et des choix nationaux. Une directive laisse plus ou moins de marge pour des choix d'opportunité.

Figure 4. Open Data : directive européenne de 2003



La mise à jour de la directive européenne de 2003[10] en élargit le champ d'application, moyennant des contraintes plus légères, aux secteurs culturels et de l'enseignement. Concernant l'eGovernment, la directive PSI 2013 :

- rend réutilisable tout contenu public pouvant être accessible légalement sur le plan national (sous réserve d'exceptions prévues au niveau des règles européennes ou nationales).
- invite les États-membres à rendre davantage de documents disponibles sous un format lisible par l'ordinateur, avec les métadonnées facilement ré-exploitables (nous reviendrons sur cet aspect au point 3 du document).

La directive insiste également sur la transparence des coûts et des licences, que nous abordons successivement dans les deux points ci-dessous. Dans le même temps, elle tente de favoriser le mouvement à travers l'organisation de workshops, de concours stimulant l'innovation et la réutilisation des données ouvertes ainsi que le financement d'actions de recherche.

[10] DE PUE E., « Open Data Challenges/Opportunities » (http://www.datadays.eu/), Ghent, 17-19 février 2014. https://docs.google.com/file/d/0B-uNSfOQ-TR_SjdZSWpOQlN4ZWM/edit?pli=1

2.2. « *Open* » ne signifie pas « *free of charge* »

Sur le plan financier, la directive permet la mise à disposition de l'information réutilisable moyennant un coût marginal (correspondant au coût de l'ouverture des données : production, diffusion…). Elle maintient à titre exceptionnel la possibilité de la mise à disposition à un coût plus élevé, le tout étant formulé de manière à laisser ouvertes certaines interprétations, ne rendant pas aisées les modalités de calcul du coût : « *Public sector bodies need to calculate charges per re-user in a way so that the total income from charging does not exceed the costs incurred to produce and disseminate the information, together with a reasonable return on investment. Public sector bodies are encouraged to apply lower charges or to apply no charges at all. On request, public sector bodies must indicate the method used to calculate charges.* »[11]

Notons que la question fait débat : certains mouvements en charge de la promotion de l'*open data* estiment que toutes les *open data* devraient être gratuites et militent dans ce sens.

Figure 5. Open data et KBO (2014)

À titre d'exemple, en Belgique, la Banque Carrefour des Entreprises (BCE)[12], source authentique des entreprises belges, a prévu, grâce à une

[11] http://ec.europa.eu/digital-agenda/en/european-legislation-reuse-public-sector-information

[12] DE SAER F., « Transparency as a driver for Opening Belgium Company Register » (http://www.datadays.eu/), Ghent, 17-19 février 2014. http://fr.slideshare.net/FrankDeSaer/open-data-vl

modification législative et un reengineering de sa base de données, à partir de mai 2014, à côté d'autres données disponibles publiquement et gratuitement (« *public search* »), la diffusion payante de certaines des données dont elle a la gestion[13]. Un enregistrement obligatoire des utilisateurs est requis (de façon à récolter leur feedback sur la qualité des données ainsi que toutes leurs questions à ce propos) et un ensemble de données sera accessible via un service web reposant sur les standards XML et SOAP (Figure 5). Ce développement vise à favoriser la mise en place d'applications en vue de renforcer la croissance économique[14] dans un secteur (les sources authentiques d'entreprises) où la diffusion payante de l'information est courante à l'étranger.

Cet exemple est particulièrement pertinent s'agissant d'une source authentique garante de la qualité de l'information diffusée. On observe un phénomène analogue dans le secteur privé bancaire, s'agissant par exemple des banques BNP Paribas Fortis et ING, lequel soulève quelques questions de vie privée (« *privacy* »).[15].

Il ne faut pas perdre de vue que dans le cas d'autres domaines d'application où la qualité, l'unicité et l'authenticité de l'information ne sont pas nécessairement prises en charge à la source, la multiplication des plateformes « *open data* » gratuites mais hétérogènes couvrant un même sujet avec la redondance d'information associée a pour conséquence une dégradation de la qualité de l'information et des coûts supplémentaires inutiles.

On observe ce phénomène non seulement sur le plan local et national, mais aussi au niveau européen où les plateformes d'*open data* sont multiples, telles que par exemple *open-data.europa.eu*, *publicdata.eu* ou encore *engagedata.eu*.

---

[13] Un sous-ensemble étant déjà à l'heure actuelle disponible moyennant paiement et signature d'une licence. http://economie.fgov.be/fr/entreprises/BCE/Entreprises/Comm/

[14] Dans l'esprit de la directive PSI 2013 : *« Public authorities produce large amounts of data that could become the raw material for new, innovative cross-border applications and services. Examples of products and services based on the re-use of public sector information (PSI) are GPS, weather forecasts, financial and insurance services. PSI is the single largest source of information in Europe. Its estimated market value is €32 billion. Re-used, this public data could generate new businesses and jobs and give consumers more choice and more value for money »*. http://ec.europa.eu/digital-agenda/en/european-legislation-reuse-public-sector-information

[15] « Fortis dévoilera vos données… mais lesquelles ? » La Libre Belgique, 19 mars 2014. http://www.lalibre.be/economie/actualite/fortis-devoilera-vos-donnees-mais-lesquelles-5328a1ea35707711f4a8c630

2.3. « Licences » et implications légales

La directive PSI 2013 renforce l'obligation de transparence sur les conditions d'accès, de licence et de coûts. Selon les enjeux de l'information et de leurs usages, les licences pourront couvrir, notamment :

- la question du *copyright* et de la propriété intellectuelle (brevets,…) ;
- les risques liés au non-respect de la vie privée, même si les données ont été anonymisées (plusieurs statistiques croisées entre elles pouvant par ailleurs contribuer à l'identification d'individus, à travers un phénomène de « triangulation ») ;
- les risques liés à l'usage inadéquat d'informations de mauvaise qualité. Ainsi le projet « *Citadel on the move* », qui encourage explicitement à publier des données ouvertes, quelle que soit la qualité de la source (« *accept mistakes : no dataset is perfect* ») se couvre avec une licence appropriée (« *Standard Open Licences such as CCBY and CC-O have watertight legal protection against liability for the accuracy of data. For more information see : https://creativecommons.org/licences* »).

Afin de déterminer la licence adéquate, on pourra en effet puiser dans les licences offertes par les creative commons, les licences les plus restrictives interdisant par exemple la transformation du contenu diffusé. Au niveau fédéral belge, le site de l'ASA/DAV mentionné plus haut propose une licence type pour la réutilisation des données du secteur public, dans la ligne de la nouvelle directive PSI 2013 et compatible avec toute licence libre qui exige a minima la mention de la paternité[16].

En effet, il faut ajouter que la multiplicité et l'hétérogénéité des licences d'un pays à l'autre ou d'une source à l'autre peuvent représenter un frein à l'utilisation des « *open data* »[17] (ainsi les « *open data* » sur une même plateforme sont parfois catégorisées selon le type de licence associé : *engagedata.eu*). Ajoutons que la problématique est particulièrement aigue dans le domaine des <u>données géospatiales et du transport</u> (information fondamentale dans la plupart des applications) : au-delà des licences, les directives légales européennes sont elles-mêmes fragmentées : à côté de la directive PSI 2013 abordée dans ce document, il existe la directive TAP-TSI (Telematics applications for Passenger Services-Technical Specifications for Interoperability

[16] http://psi.belgium.be/sites/default/files/assets/Licence_20131014_FR.pdf
http://psi.belgium.be/sites/default/files/assets/Licence_20131014_NL.pdf

[17] JANSSEN K.,« One licence to rule them all: Very ambitious or just plain delusional?» (http://www.datadays.eu/), Ghent, 17-19 février 2014 https://docs.google.com/file/d/0B-uNSfOQ-TR_WXRacDNVcUc1UFE/edit?pli=1

Regulation (EU) N° 454/2011), INSPIRE (Infrastructure for Spatial Information in the European Community N°2007/2/EC) et Rail Passenger Rights Regulation (EC) N° 1371/2007[18].

3. Typologie des « *open data* » dans l'eGovernment

Dans la ligne de la définition présentée au point 1, les *open data* sont a priori destinées à être ouvertes à tous (3.1.). Nous verrons toutefois que certaines d'entre elles, dans des circonstances spécifiques à l'eGovernment, peuvent être plus ou moins contrôlées (3.2) ou, en marge de la directive européenne, restreintes à un sous-ensemble donné d'utilisateurs, « open » ne signifiant pas nécessairement « public » (3.2).

3.1. Open data ouvertes et publiques

3.1.1. Statistiques et ranking

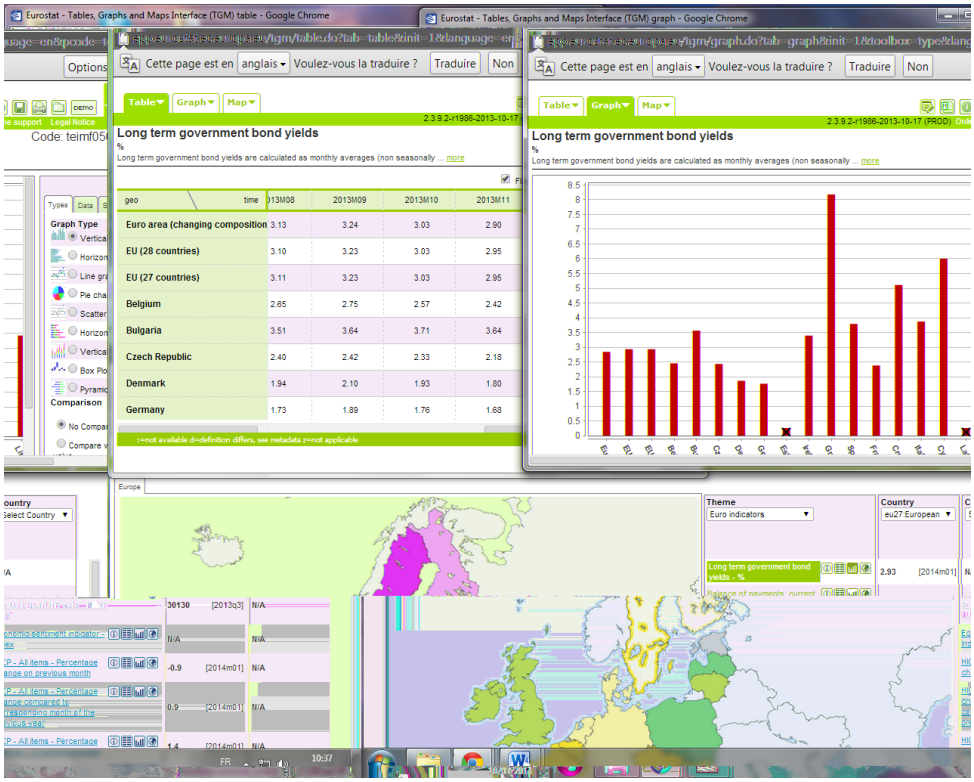
Parmi les « *open data* » ouvertes et publiques, on distingue les données statistiques, qui présentent l'avantage d'être explicitement publiées de manière discrète (par opposition aux données continues de type « *real time* »), ponctuellement statique et d'offrir des outils d'interprétation utiles si elles sont accompagnées de métadonnées suffisamment complètes quant à leur méthode de conception, leur périodicité, leurs sources, leur couverture spatio-temporelle, etc. (ce qui ne signifie pas qu'elles ne sont jamais affectées par des problèmes de qualité ni que leur exploitation est aisée, mais ce point mériterait une étude à part entière).

Eurostat[19], pour la commission européenne, présente un exemple type de données statistiques dont la production est contrôlée à la source et dont la diffusion est ouverte. Celles-ci sont particulièrement bien documentées : la plateforme inclut des outils de visualisation et d'extraction pour téléchargement automatisé (Figure 6). Eurostat représente par ailleurs une source faisant autorité dans le contexte de la multiplication des plateformes d'*open data* évoquées plus haut.

[18] SZELIGOWSKA D.: « *Extending access to travel and traffic data"*» (http://www.datadays.eu/), Ghent, 17-19 février 2014. https://docs.google.com/file/d/0B-uNSfOQ-TR_ZTFlcDU2WDlGcEk/edit?pli=1

[19] http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home

Figure 6. Portail statistique d'Eurostat



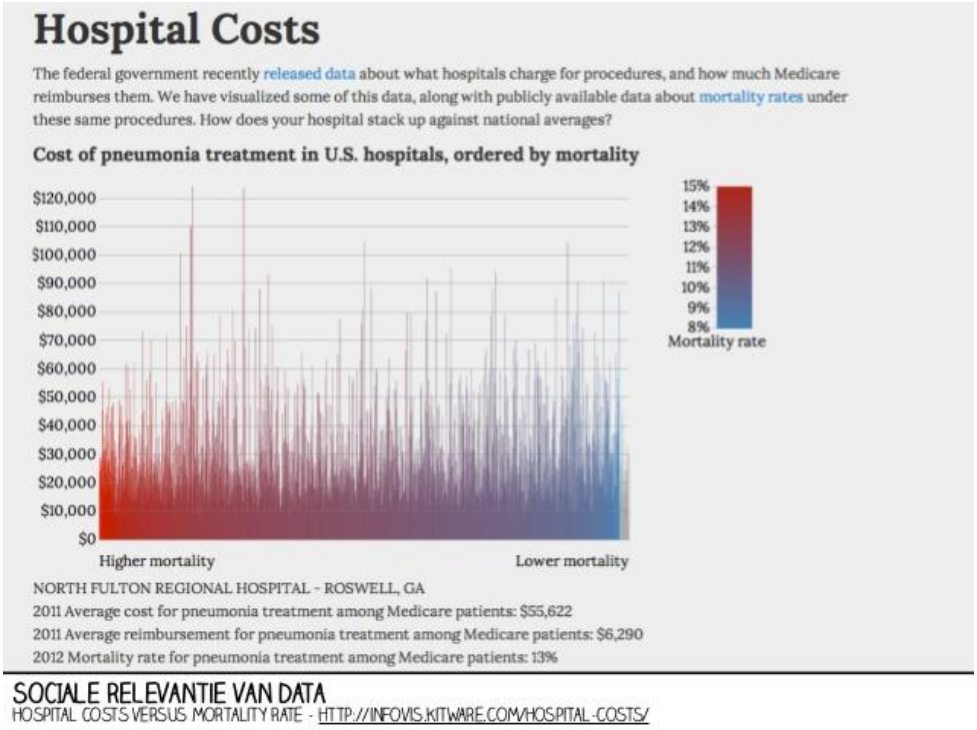
Dans certains cas, les statistiques sont produites afin de mettre des services en concurrence et d'en améliorer la qualité et la transparence : ainsi *l'Agence fédérale pour la sécurité de la chaîne alimentaire* (AFSCA) envisage-t-elle la publication de rapports synthétiques de contrôle d'hygiène[20].

Citons également l'application suivante (Figure 7) développée aux États-Unis et fournissant par hôpital les coûts des soins de certaines maladies eu égard au taux de mortalité observé l'année suivante[21].

[20] DE PUE E., « Open Data Challenges/Opportunities » (http://www.datadays.eu/), Ghent, 17-19 février 2014. https://docs.google.com/file/d/0B-uNSfOQ-TR_SjdZSWpOQlN4ZWM/edit?pli=1

[21] VANDE MOERE A., « Information visualization : from analysis to the communication of data insights » (http://www.datadays.eu/), Ghent, 17-19 février 2014. http://www.slideshare.net/infoscape/information-visualization-analysis-and-communication-of-insights

Figure 7. Information visualization : from analysis to the communication of data insights *(VANDE MOERE A)*



3.1.2. Données « *real time* » et données de gestion

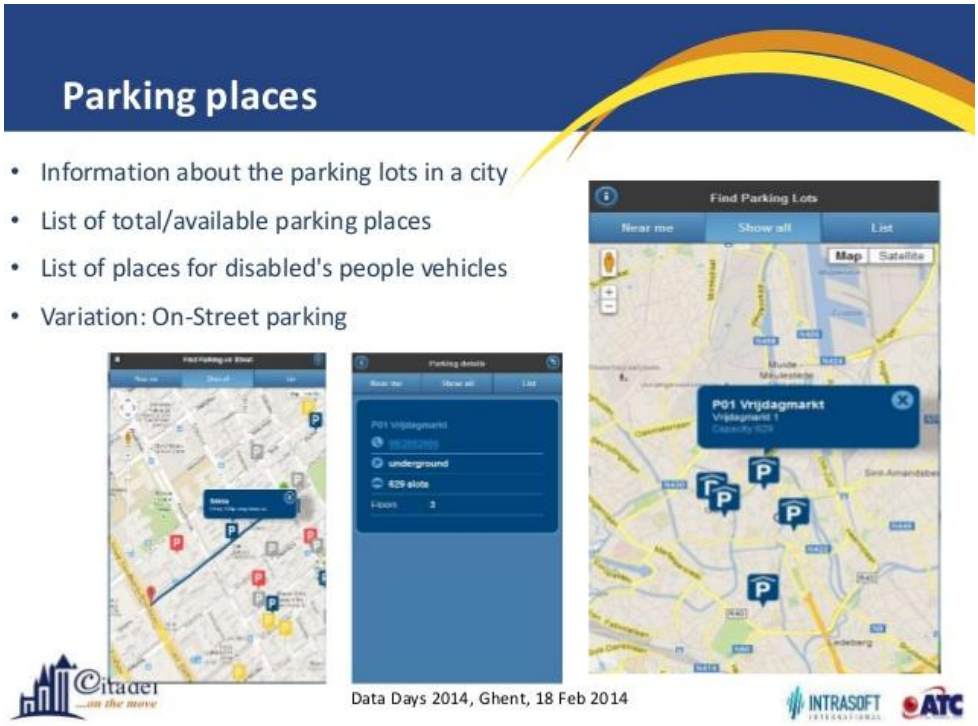
La difficulté que soulèvent les données de type « real time » ou même de gestion, lorsqu'elles sont sujettes à évolution dans le temps, c'est qu'elles peuvent rapidement devenir obsolètes, de qualité inadéquate et impropres à l'usage. Aussi leur maintenance et leur gestion nécessitent-elles une intervention contrôlée en amont si l'on veut éviter ces écueils, dont les ressources peuvent justifier un service payant en aval, comme l'offre la BCE. Maintenant, d'autres initiatives telles que le projet Citadel des « *smart cities* » en Flandre proposent d'emblée la diffusion de données ouvertes, quelle que soit leur qualité (même douteuse), de façon à encourager le mouvement « open » gratuit dans certaines communautés.

Le choix entre l'une ou l'autre stratégie repose sur des décisions de gestion, en fonction du contexte et en connaissance de cause, sachant les problèmes de qualité que soulève déjà à l'heure actuelle la diffusion massive d'*open data* non contrôlées, obsolètes et incomplètes sur Internet à travers le monde, telles que les décrit Gartner dans son rapport de novembre 2013, cité dans l'introduction générale (point 1).

Un cas d'application très fréquent concerne les « *smart cities* » et la publication de données « *real time* » les concernant impliquant le trafic, la mobilité, les mouvements de foule…

Par exemple, les places de parking disponibles un jour de marché peuvent être diffusées en « real time » : projet flamand Citadel (Figure 8)[22].

Figure 8. Application issue du projet Citadel



Le suivi de l'impact des catastrophes naturelles, telles que l'ouragan Sandy, incluant des données publiques météorologiques mais aussi les constatations des citoyens constitue un autre exemple récent d'*open data* « *real time* » de plus grande ampleur (Figure 9).[23]

[22] « Open. Convert. Use. The Citadel Open Data Ecosystem ». (http://www.datadays.eu/), Ghent, 17-19 février 2014 http://fr.slideshare.net/thimothoeye/citadel-technical

[23] HOWARD A., « Tracking the data storm around Hurricane Sandy. When natural disasters loom, public open government data feeds become critical infrastructure », http://strata.oreilly.com/2012/10/real-time-data-storm-in-hurricane-sandy-open-data.html

Figure 9. Suivi « real time » de l'ouragan Sandy

3.2. Degré de contrôle à la source des « *open data* »

La perfection n'existe pas et ce, même dans le cas des sources authentiques, dont la définition légale est pragmatique et sectorielle. La notion de « source authentique » renvoie à un service responsable du contrôle et du suivi dans le temps de la qualité de l'information produite[24]. Toutefois, on peut faire une distinction entre les « *open data* » :

- dont la production est contrôlée par une source authentique (BCE, Eurostat, PubChem…) ;
- dont la production est libre à la saisie, potentiellement ouverte à tous, faisant au mieux l'objet d'un contrôle « ex post » via un DQ Tool, comme sur la plateforme www.engage.eu.

Il est évident que la qualité et la fiabilité de l'information seront a priori meilleures dans le premier cas, l'utilisateur ayant toujours la possibilité de de s'adresser au producteur de l'information pour toute question sémantique ou relative à la mise à jour des données.

3.3. « *Closed data* »

A côté de la directive PSI 2013, dans l'intérêt commun opérationnel d'un domaine d'application donné, des informations peuvent être conçues de façon à être ouvertes et accessibles dans le cadre d'un

[24] Pour plus de details, voir : BOYDENS I., HULSTAERT A. et VAN DROMME D., « Gestion intégrée des anomalies. Evaluer et améliorer la qualité des données » rapport de recherche, Bruxelles, Smals, del11-trim1-04, mars 2011, p. 17-19. http://www.smalsresearch.be/download/research_reports/deliverable/Gestion%20intégrée%20des%20anomalies.pdf

réseau secondaire, ciblé, et sécurisé. On parlera « pragmatiquement » (et non juridiquement) de « *closed data* » interopérables

4. Aspects fonctionnels

Nous abordons ici successivement les éléments fonctionnels à prendre en considération en vue de la diffusion d'*open data* : les formats et standards existants et conseillés, le recours aux « data quality tools », le recours à des outils de conversion et de gestion du multilinguisme, aux standards d'interopérabilité sémantique et, enfin, les plateformes de mise à disposition des *open data* donnant naissance à des applications de « *citizen engagement* ».

4.1. Formats de données et de métadonnées

Dans le cadre du Web sémantique, une échelle de qualité des données ouvertes a été proposée en 2010 par T. Berners-Lee (tableau 1). On y a fréquemment recours pour situer la qualité d'une donnée ouverte[25].

Tableau 1. Échelle de qualité des formats de données (T. Berners-Lee)

★	Données non filtrées (éventuellement dégradées), par exemple mises en ligne avec n'importe quel format
★ ★	Données disponibles sous forme structurée (ex : données tabulaires en CSV, XML, Excel, RDF)
★ ★ ★	Données librement exploitables - juridiquement (cf. licences) et - techniquement (dans des formats ouverts non propriétaires, pas sous Excel notamment)
★ ★ ★ ★	Données identifiées par des URL (avec date de mise à jour) afin que l'on puisse « pointer » un lien vers elles (et les retrouver éventuellement mises à jour)
★ ★ ★ ★ ★	Données liées à d'autres données (« *linked data* ») pour les contextualiser et les enrichir

Cette échelle concerne les données numériques de base, mais elle peut être adaptée à des informations plus complexes (photos, vidéos, rapports, études, etc.). Pour plus d'informations, nous renvoyons à la source correspondante du W3C : http://www.w3.org/standards/

Dans la pratique, le niveau 3 (données sous licence et sous un format ouvert), s'il est accompagné d'informations sémantiques (date de mise à jour, définition…) est un minimum acceptable. Le niveau 5, par contre, s'il repose sur des standards matures, demande, en raison de sa complexité, un savoir-faire en termes de mise à jour potentiellement coûteux. Il n'est dès lors recommandé dans les projets incluant un nombre important de données stratégiques évolutives et transactionnelles que si l'on dispose

[25] http://fr.wikipedia.org/wiki/Donn%C3%A9es_ouvertes

d'un budget et de ressources importants en termes de savoir-faire technique et de main-d'œuvre pour la gouvernance et la maintenance.

Naturellement, les données doivent être documentées pour être compréhensibles et ces standards s'appliquent également aux formats des méta-informations correspondantes, lesquelles doivent idéalement faire l'objet d'une gestion rigoureuse[26].

4.2. Data Quality tools

Comme mentionné dans l'introduction générale, la mise en œuvre sur Internet des normes du Web sémantique, dans un environnement ouvert et non contrôlé, a vite posé des inévitables problèmes de qualité de données, s'agissant des données ouvertes. Nous avons synthétisé ce phénomène en 2011, prévisible dès l'émergence du web sémantique dans un blog intitulé « *Linked open data quality around the* clock »[27].

La question reste d'actualité (et le problème ne cesse dès lors de s'amplifier) lorsque les données ne sont pas contrôlées. Afin de traiter cette problématique a posteriori, le projet européen « Engage » (www.engage.eu) a par exemple recours au Data Quality Tool libre « Open Refine »[28]. À côté des opérations de data profiling, il est fortement recommandé de maintenir une gestion des versions des données.

4.3. Outils de conversion

Certains projets, tels que le projet Citadel déjà cité à propos des « *smart cities* », considèrent le format csv comme le « couteau suisse » de l'*open data* en raison de sa facilité d'usage pour les non-informaticiens (même si ce couteau coupe parfois mal, en l'absence de sémantique). Par ailleurs, ce format n'est a priori pas très riche. Aussi la plateforme « *Citadel* » offre-t-elle des outils de conversion en vue de concevoir des services web interactifs destinés aux applications mobiles et reposant potentiellement sur des données géospatiales[29].

---

[26] A propos des méta-informations, voir : http://www.ulb.ac.be/cours/iboydens/e-gouvernement.pdf

[27] http://www.smalsresearch.be/archives/2048

[28] VAN HOOLAND S. et VERBORGH R., « Linked data for libraries, archives and museums. How to clean, link and publish your metadata ». Brimingham-Mumbai : Facet Publishing (to be published in 2014). DE WILDE M. et VERBORGH R., « Using OpenRefine ». Brimingham-Mumbai : Packt Publishing, 2013. Voir aussi à propos de cet outil la « quick review » publiée en 2013 par V. BERTEN : http://www.smalsresearch.be/publications/quick-reviews.

[29] « Open. Convert. Use. The Citadel Open Data Ecosystem ». (http://www.datadays.eu/), Ghent, 17-19 février 2014. http://fr.slideshare.net/thimothoeye/citadel-technical

4.4. Outils de gestion du multilinguisme

Dans le contexte de l'*open data* et du *big data*, on est naturellement confronté au multilinguisme, même si l'anglais prédomine. Des techniques de NLP (*Natural Language Processing*) peuvent alors être mobilisées[30].

4.5. Standards d'interopérabilité sémantique

Afin d'obtenir des résultats opérationnels dans le contexte d'applications multisources, des séries hétérogènes d'*open data* doivent être croisées[31]. À cette fin, des standards d'interopérabilité sémantique[32] ainsi qu'un langage de requête (SPARQL) sont proposés par le W3C[33]. À ce propos, nous rappelons les réserves émises au point 4.1 ci-dessus concernant les formats et standards. Dans le domaine de l'eGovernment, il existe des langages d'interopérabilité sémantique en cours de développement en vue d'être adaptés aux « *open data* » sur le plan local (par exemple, OSLO, *Open Standard for Local Authorities - Open Standaard voor Lokale Overheden in Vlaanderen*) et européen (par exemple, ISA, *Interoperability Solutions for European Public Administrations*[34]). Des travaux sont menés en vue de leur développement cohérent[35].

4.6. Plateformes de « *citizen engagement* »

Progressivement associées dans le cadre de plateformes adéquates, les *open data* ont pour finalité, sur le plan applicatif, de donner le jour à des *mashups* et, par exemple, à des services de « *citizen engagement* ». Citons à titre d'exemple la plateforme européenne suivante qui propose à la fois des *open data* documentées et plusieurs exemples d'applications : http://publicdata.eu/related

---

[30] VAN HOOLAND S., DE WILDE M., VERBORGH R., STEINER T. et VAN DE WALLE R., « Exploring Entity Recognition and Disambiguation for Cultural Heritage Collections ». In Literary and Linguistic Computing (accepted for publication).

[31] VAN HOOLAND, S. et VERBORGH, R. « Joining the Linked Data Cloud in a Cost-Effective Manner. ». Information Standards Quarterly (ISQ), Spring/summer 2012, v.24, p. 24-29, 2012

[32] Dans le contexte de l'eGovernment, voir le blog suivant : BOYDENS I. « L'intéropérabilité sémantique : une révolution ? Les normes SKOS (W3C, 2009) et ISO 25964 », 10/04/2012. http://www.smalsresearch.be/archives/4091

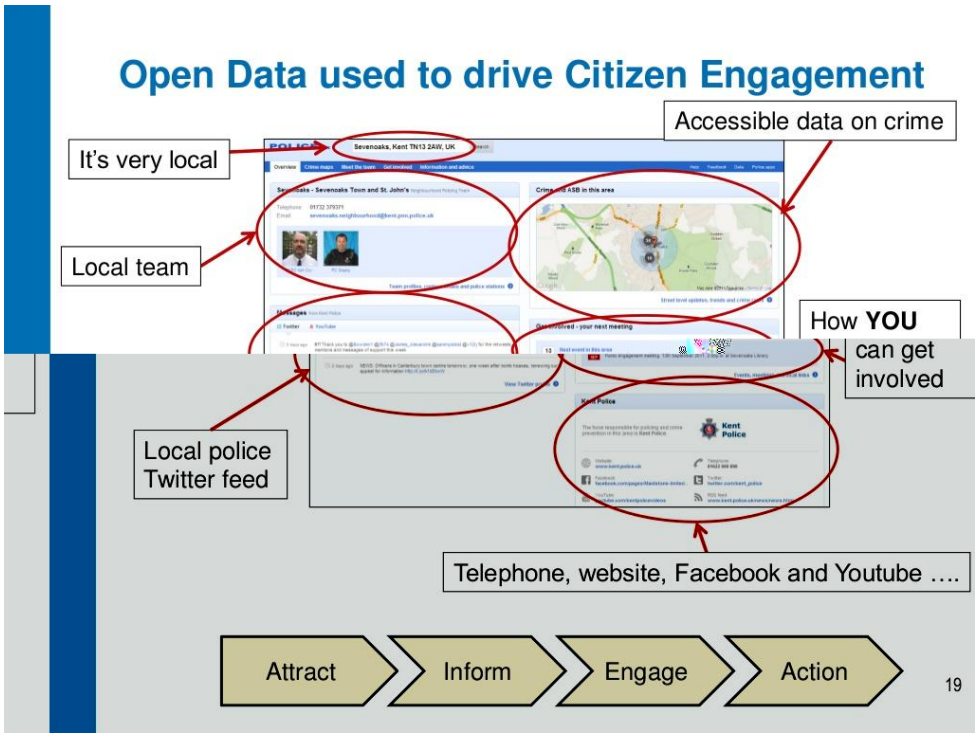
[33] Pour plus d'information, voir le site du W3C : http://www.w3.org/TR/rdf-sparql-query/

[34] http://ec.europa.eu/isa/

[35] BUYLE R., « Extending the EU ISA Core Vocabularies to a local level by the OSLO ». (http://www.datadays.eu/), Ghent, 17-19 février 2014. http://fr.slideshare.net/raf_b/oslo-how-semantic-collaboration-is-going-to-be-a-cornerstone-of-the-next-generation-smart-cities

Parmi les applications fréquemment évoquées, sur la base de l'expérience de la plateforme britannique d'*open data*[36], où le grand public s'engage tendanciellement plus volontiers, citons : les données de proximité et locales et au sein de celles-ci : l'aide à la résolution de crimes (Figure 10)[37], les informations concernant la santé et les hôpitaux, l'aide aux enfants (écoles, crèches…), l'aide aux personnes âgées ou encore l'intervention de bénévoles lors de catastrophes naturelles. Un autre type d'application est « l'*eRulemaking* », qui encourage les citoyens à participer au débat public en vue de la mise en place de nouvelles réglementations. Il est reconnu que le « *return on investment* » de telles approches est toutefois difficile à calculer et encore plus à prévoir.

Figure 10. Plateforme d'aide à la résolution de délits



Une question de fond consiste à savoir s'il faut d'abord diffuser des « *open data* » en ignorant les usages a priori imprévisibles auxquels elles pourront donner lieu (c'est l'esprit de l'initiative http://dbpedia.org/) ou, à l'inverse, d'abord concevoir de potentielles « *killer applications* » et ensuite, mobiliser les données destinées à les alimenter.

[36] STOTT A., « Getting Open Data Used » (http://www.datadays.eu/), Ghent, 17-19 février 2014. http://www.slideshare.net/dirdigeng/getting-open-data-used

[37] Cette application est également évoquée s'agissant de la police de Chicago dans le blog suivant : BOYDENS I., « Mapping the World of Data Problems" : la qualité des données vue par la communauté IT » , 03/04/2013 http://www.smalsresearch.be/archives/5398#sthash.558KLtrg.dpuf

5. Conclusions et recommandations

La mise à jour récente de la directive européenne PSI en 2013 couplée au phénomène croissant de l'*open data* demandait qu'une note de recherche sur la question en présente un aperçu critique.

Sur le plan légal, toutefois, nous n'avons fait qu'esquisser les grandes lignes des évolutions à venir et des juristes devront être mobilisés notamment pour approfondir et préciser la directive PSI 2013 dans son ensemble ainsi que les questions de licences et de calcul de coût (point 2). En Belgique, des organismes tels que Fedict ou l'ASA-DAV suivent ces questions de près.

Au point 4, nous avons, sur la base d'une typologie des *open data* (point 3), passé en revue les étapes fonctionnelles à prendre en considération en vue de leur mise en place et nous avons émis à plusieurs reprises des conseils ou avis en vue de la conception des *open data* ou de leur interprétation, selon les circonstances.

Plusieurs bonnes pratiques et recommandations complémentaires peuvent par ailleurs être formulées :

- Avant de penser à la production de nouvelles plateformes de « données ouvertes », il peut être utile de recenser et de cartographier toutes les données publiques déjà mises à disposition sur les portails et sites de l'eGovernment, d'en vérifier la qualité, l'usage, la conformité aux licences… et d'en mettre en valeur, si le besoin s'en fait sentir, le caractère ouvert. Ceci sera d'autant plus pertinent à la faveur d'un reengineering en cours. De nombreuses données ouvertes existantes ne sont en effet pas suffisamment connues ou ne se trouvent pas sous un format ou une licence adéquats. Certaines d'entre elles ne sont pas utilisées et sont laissées en friche, leur qualité se déteriorant inévitablement avec le temps (ce qu'illustre l'adage « *use it or lose it* »). On pourra également découvrir à l'occasion de cet inventaire de nombreuses données « ouvertes » redondantes et mises à jour à des rythmes différents, dont l'exploitation demande a posteriori de nouvelles opérations de tests et de correction (mécanisme qu'illustre le concept bien connu de « *ghost factory* », désignant le temps et l'argent consacrés inutilement à la production de problèmes et à leur correction au sein d'une même entité). Un inventaire continu de cet ensemble permettra d'en améliorer et d'en maintenir la qualité dans le temps.
- La qualité des données ouvertes dans le cadre de l'eGovernment[38] est un point crucial : plus les informations seront documentées, contrôlées et régulièrement mises à jour, plus les applications qui les exploiteront seront efficaces et plus l'on pourra ralentir l'hémorragie actuelle d'*open data* obsolètes et de piètre qualité diffusées sur le Web. La gestion des versions des données

[38] BOYDENS I., « Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium ». In ASSAR S., BOUGHZALA I. et BOYDENS I., éds., « Practical Studies in E-Government : Best Practices from Around the World » , New York, Springer, 2011, p. 113-130 (chapitre 7).

publiées est en particulier un élément fondamental. La « qualité totale » n'existe pas, même dans le cas des « sources authentiques », mais du moins, dans ce cas, des responsabilités claires sont établies en vue d'en assurer le suivi continu dans le temps eu égard aux usages.

- Le phénomène des *open data* peut à cet égard être exploité en vue d'améliorer la qualité de sources mises à la disposition du public via un système de feedback (tel que celui proposé par la BCE), dans le cadre d'une « gestion collaborative des anomalies » par plusieurs communautés d'utilisateurs potentiellement intéressés par les mêmes données authentiques.
- Une approche complémentaire du même ordre peut être menée dans des secteurs où le ROI de l'approche « *data quality* » doit être maximisé et où l'on n'est pas soumis à la contrainte juridique de la force probante, les commentaires librement fournis par les utilisateurs faisant l'objet d'une interprétation statistique en vue de permettre d'améliorer de manière semi-automatique la qualité des métadonnées. Un projet de ce type a été développé avec succès pour le *National Archives of the Netherlands* aux Pays-Bas et le *September 11th Memorial and Museum* de New York[39].
- Il est utile de rappeler, même si cela relève du bon sens, les principes de gouvernance et de bonnes pratiques conceptuelles et fonctionnelles de base lors de la conception de toute nouvelle application informatique[40]. Outre l'avantage immédiat, les informations correspondantes seront d'autant plus interopérables dans le cadre d'une application de type « *open government* ».

Les systèmes d'information empiriques se transforment avec l'interprétation des valeurs qu'ils permettent d'appréhender. Leurs enjeux sont stratégiques dès lors qu'ils sont des instruments d'action sur le réel[41] . La mise en place de la directive européenne PSI 2013 et la question des *open data* dans l'eGovernment soulèvent à cet égard des enjeux juridiques, financiers, éthiques, opérationnels et sémantiques stratégiques concernant la gouvernance et les usages de l'information administrative. En vue d'en assurer le suivi, une coordination des initiatives « *open data* » dans le secteur de l'eGovernment rassemblant des experts et des décideurs semble indispensable.

---

[39] BOYDENS I. et VAN HOOLAND S., « Hermeneutics applied to the quality of empirical databases ». In Journal of documentation, volume 67, issue 2, 2011, pp. 279-289.

[40] BOYDENS I., « E-gouvernement en Belgique. Un retour riche d'expériences ». In L'informatique professionnelle (Dossier spécial « Services publics »). Paris : Editions Gartner France, Numéro 217, octobre 2003, p. 29-35. http://www.ulb.ac.be/cours/iboydens/e-gouvernement.pdf

[41] BOYDENS I., « L'océan des données et le canal des normes ». In CARRIEU-COSTA M.-J., BRYDEN A. et COUVEINHES P. éds, Les Annales des Mines, Série Responsabilité et Environnement (numéro thématique : « La normalisation : principes, histoire, évolutions et perspectives »), Paris, n° 67, juillet 2012, pp. 22-29. http://www.ulb.ac.be/cours/iboydens/annales.pdf

Le centre de compétences Data Quality fait partie de la section Recherche de Smals. Le centre de compétences peut se targuer d'une **expérience intensive sur le terrain depuis 2004**. Pour la plupart des projets, les membres de la cellule Data Quality travaillent main dans la main avec diverses divisions de Smals, comme la section Développement des applications & Projets, Traitement de l'information ainsi que la section Statistiques ou avec les services de clients et d'institutions membres. Les différentes tâches sont ensuite réparties en concertation avec chacun.

En parallèle avec les missions de consultance autour de la qualité des bases de données administratives des institutions membres, les collaborateurs du centre de compétences donnent aussi des formations et mènent des recherches actives dans ce domaine.

Voir le site web de Smals: https://www.smals.be/fr/content/data-quality

Het competentiecentrum Data Quality maakt deel uit van de sectie Onderzoek van Smals. Het competentiecentrum heeft een **intensieve ervaring op het terrein sinds 2004**. De leden van de cel Data Quality werken voor de meeste projecten samen met diverse afdelingen van Smals, zoals de sectie Toepassingsontwikkeling & Projecten, Informatieverwerking en de sectie Statistieken, of met diensten van klanten en lidinstellingen. De verschillende taken worden dan in onderling overleg verdeeld.

Parallel met de consultancyopdrachten omtrent de kwaliteit van de administratieve databases van de lidinstellingen geven de medewerkers van het competentiecentrum ook opleidingen en verrichten zij actief onderzoek in dit domein.

Zie website Smals: https://www.smals.be/nl/content/data-quality

La section Recherche de Smals produit régulièrement des publications couvrant de nombreux domaines du marché IT actuel. Vous pouvez obtenir ces publications via le site web de la section Recherche :

http://www.smalsresearch.be