Apports des données publiques aux SI de Santé : étude exploratoire et premiers résultats

Gabriella Salzano[*,**], **Joumana Boustany**[*,***]

[*] *Équipe de recherche DICEN (Dispositifs d'information et de communication à l'ère numérique) EA 4420, CNAM. Case I461 -2, rue de Conté – 75003 Paris*

[**] *Université Paris Est Marne-la-Vallée*
5, Bd. Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée Cedex 2
gabriella.salzano@univ-mlv.fr

[***] *Université Paris Descartes – IUT, 143, avenue de Versailles - 75016 Paris*
joumana.boustany@parisdescartes.fr

RÉSUMÉ. *Ce papier s'inscrit dans la thématique de l'Ingénierie des Systèmes d'Information de Santé (SIS). Dans un contexte favorable à l'ouverture des données publiques (DP), cette recherche étudie les opportunités offertes par ces données aux SIS. Produites par des administrations dans le cadre de leurs missions, les DP sont caractérisées par leur fiabilité et par la possibilité de réutilisation. À l'aide d'une démarche empirique, nous réalisons une macroclassification des SIS basée sur des critères fonctionnels et de couverture territoriale. Notre analyse de l'offre actuelle des DP en santé, diffusées par les majeures plateformes gouvernementales et transnationales, s'appuie sur plusieurs critères (couverture thématique et territoriale, volumes, formats, organisation). Cette offre est corrélée aux types de SIS. Par ailleurs, nous identifions les principaux freins qu'il convient d'éliminer et les tendances globales.*

ABSTRACT. *This paper concerns the theme of Engineering Information Systems of Health (HIS). In a context that favors open public data (PD), this research analyzes the opportunities provided by these data to HIS. By being provided by governments as part of their missions, DP are characterized by reliability and reusability. Following an empirical approach, we perform a HIS macro classification, based on functional criteria and territorial coverage. We analyze the current supply of health PD issued by the major governmental and transnational platforms, using several criteria (thematic and territorial coverage, volume, format, organization). We correlate this offer to different types of SIS. Finally, we identify the main barriers to be eliminated and the overall trends.*

MOTS-CLÉS : *Systèmes d'Information de Santé, ingénierie, données publiques, qualité des données, fiabilité, réutilisation.*

KEYWORDS: *Health Information Systems, engineering, public data, data quality, reliability, reuse.*

1. Introduction

Les systèmes de santé, comme définis par l'OMS (Organisation mondiale de la santé), assurent des prestations de services qui améliorent, maintiennent ou rétablissent la santé des individus et des communautés. Ces systèmes doivent veiller aussi à améliorer les conditions sociales, économiques ou environnementales des populations.[1]

Les Systèmes d'Information de Santé (SIS) doivent faire face à la complexité de la santé publique, de la gestion des risques industriels et environnementaux, des prestations de soins et des services de santé. Des structures collaboratives et transversales, comme des systèmes de télémédecine, d'hospitalisation à domicile ou plus généralement des réseaux de santé, se mettent en place progressivement pour apporter des services de santé plus performants.

Dans ce contexte, cette recherche concerne les besoins d'évolution des SIS et focalise sur la mise à disposition et la fiabilité des données. Ces critères caractérisent les données publiques (DP) qui font l'objet d'une ouverture et d'une possibilité de réutilisation dans un cadre réglementaire favorable. C'est le cas de l'Union européenne, des États-Unis, du Royaume-Uni, de la Nouvelle-Zélande, de la France, etc. Les DP sont au cœur de l'innovation. C'est ainsi que les États-Unis ont décerné le prix de l'innovation 2011 à l'application mobile Asthmapolis, réalisée en partenariat avec les instances publiques : les données recueillies par des patients asthmatiques équipés d'inhalateurs géolocalisés sont consolidées et agrégées ; elles permettent à des médecins, scientifiques et organismes de santé publique d'identifier des déclencheurs environnementaux, surveiller les émissions de polluants, changer les lois de zonages, donc améliorer globalement la gestion de l'asthme.

Ce papier présente une étude exploratoire sur les apports des DP aux évolutions des SIS en termes d'enrichissement par de nouvelles données. Nous visons à fournir des éléments de réponse aux questions suivantes : (Q1) peut-on corréler les besoins en évolution des SIS avec les DP ? (Q2) quelle est la disponibilité actuelle de DP en santé et quelles sont les caractéristiques principales de ces données ? (Q3) quels sont globalement les apports des DP à l'évolution des SIS et quels sont les principaux freins qui restent à lever ?

Une *démarche* empirique conduit aux étapes suivantes :

– Synthèse des définitions et des approches concernant l'ouverture des données et les données publiques (§2).

– Analyse structurée des besoins d'évolutivité des SIS et mise en corrélation avec la diffusion des DP (§3).

– Sélection d'un panel représentatif de plateformes gouvernementales, transnationales et nationales, publiant des DP en santé (§4).

---

[1] World Health System: http://goo.gl/4ckpi

– Évaluation globale des apports actuels des DP aux besoins d'évolution des différents types de SIS ainsi que des obstacles qui subsistent (§5).

2. État de l'art synthétique

Cette recherche s'inscrit dans le thème de l'« ingénierie des évolutions » des SI, qui comprend l'analyse et la conception de nouveaux SI devant faire face aux évolutions des organisations, des données et des processus métiers [Akoka, Comyn-Wattiau, 2009]. Elle focalise sur la dynamique engendrée par la diffusion des données publiques sous l'impulsion du mouvement d'ouverture des données.

L'ouverture, principe maître de l'évolution des SI, concerne aussi bien les données que les standards, les logiciels, les licences et les usages [Shadbolt, 2010]. Un jeu de données est considéré comme ouvert s'il satisfait les critères formulés par l'Open Knowledge Foundation (http://goo.gl/JKiFr), notamment l'accessibilité, l'absence de restrictions techniques et la réutilisation.

En France, la définition des données publiques s'appuie sur la loi du 17 juillet 1978. Il s'agit de données collectées, maintenues et utilisées par les organismes publics dans le cadre de leurs missions a priori gratuites, mais ceci n'est pas toujours le cas.

Les DP mises à disposition par des plateformes de diffusion, désignées par la suite par *DPPD (Données Publiques issues de Plateformes de Diffusion)*, sont intéressantes par leur ouverture et fiabilité.

Plusieurs plateformes, comme data.gov.uk, prévoient des mécanismes d'évaluation et d'enrichissement collectif des données (annotations de la part des utilisateurs). Dans ce cas, les écarts entre la qualité des données attendues et celles des données produites devraient s'estomper progressivement [ENPC, 2011]. Une analyse des formats les plus répandus des DP dégage des freins et des tendances à l'interopérabilité de ces plateformes [Boustany, Salzano, 2011].

3 Analyse structurée des besoins d'évolutivité des SIS (2 pages)

Plusieurs typologies des SIS ont été proposées : [Wager, Lee, Glaser, 2009] classifient les SIS au sein de l'hôpital, [Chekli, Lejeune, 2009] proposent une classification globale basée sur des critères de contrôle. Nous classifions les SIS en quatre catégories définies selon deux critères : la couverture territoriale et le type d'objectifs prioritaires poursuivis. Les niveaux de couverture territoriale ne sont pas étanches : en effet, « du fait de sa dimension stratégique, l'interopérabilité doit être conçue dans un contexte international, organisée dans le cadre d'une politique nationale et mise en œuvre au niveau local » [Fieschi, 2011].

Nous présentons ci-dessous ces classes et leurs corrélations possibles avec les DP.

1) couverture territoriale locale et vocation fortement opérationnelle. SI cliniques et administratifs ou « ancillaires » (SI de radiologie, pharmacie...). Utilisés et contrôlés par les prestataires de soins et de services de santé, dans différents contextes : en individuel, en hôpital ou au sein de centres médicaux ou cliniques.

2) couverture territoriale étendue (départementale, régionale) et vocation opérationnelle. SI construisant des dossiers de santé longitudinaux, comme le Dossier Médical Personnel (réseaux de santé ou d'hospitalisation à domicile), SI dits d'eHealth (SIS de télémédecine, de télémonitoring ou centres d'appel). Les premiers doivent assurer la continuité et la coordination de soins inter-épisodes, tout en maîtrisant des coûts à l'échelle d'un territoire sanitaire élargi. Les deuxièmes exploitent largement le transfert électronique de données numériques, vidéo ou multimédia, au niveau départemental ou régional.

3) couverture territoriale nationale et vocation décisionnelle. SI de santé publique, d'épidémiologie, SI de simulation, de formation à distance. Ils sont généralistes ou focalisés sur une thématique particulière : évaluer des prestations de soins et des services de santé, l'organisation des structures de soin, la qualité (performance, efficacité, respect de protocoles, coûts), simuler des problèmes cliniques (influence de facteurs génétiques sur les pathologies), d'organisation (localisation d'unités de soin) ou régulation (impact des règlementations).

4) *couverture territoriale internationale et vocation décisionnelle.* SIS sous la responsabilité d'organismes internationaux promouvant la qualité des politiques publiques de santé, comme le SIS de l'OMS. Ils consolident et/ou agrègent des données issues de SIS nationaux, pour établir des indicateurs de santé qui facilitent des études comparatives.

Corrélations potentielles entre les classes de SIS et les DP.

Classe 1) En l'état actuel, ces systèmes n'alimentent pas directement des plateformes publiques ; les données personnelles et nominatives étant exclues des plateformes de diffusion des DP. Ils transmettent des résultats d'examens d'anatomie pathologique aux registres nationaux de cancers et profiteraient de services d'aide à la prescription ou au diagnostic.

Classe 2) Ces SIS s'appuient sur les SIS de classe 1 et nécessitent aussi de données extra-médicales. Elles concernent par exemple les structures sociales, les services de transport ou les modes de prise en charge des soins sur les territoires concernés.

Classe 3) La conception de ces SIS nécessite des données facilement identifiables, rapidement disponibles, gratuites, réutilisables, exhaustives, définies à plusieurs échelles spatio-temporelles. Identifier ces sources est une activité particulièrement coûteuse en temps et ressources [Salzano, 2008], [Affset, 2008]. Il est difficile de suivre des indicateurs de santé publique, surveiller des inégalités sociales et territoriales au niveau national et réaliser des comparaisons à l'échelle européenne et internationale [HESP, 2009]. Disponibilité et qualité de données nécessitent souvent des compromis

ainsi que des retours vers les fournisseurs de données. Sur des territoires de petite taille ou de faible densité, les données sur les causes de décès sont insuffisantes et leur couverture spatio-temporelle est hétérogène. L'hétérogénéité des classifications est un frein à l'hétérogénéité sémantique de ces systèmes, qui rejaillit au niveau international [Fieschi, 2011].

Classe 4) Les principaux freins aux développements de ces SIS sont la faible qualité des données communiquées par les états (disponibilité et fraîcheur), l'hétérogénéité des classifications de santé utilisées et l'identification, de façon universelle et non ambiguë, de plusieurs concepts, comme les structures de santé (organisations, services, unités fonctionnelles, etc.) ou les systèmes techniques nécessaires à dispenser des soins.

Cette analyse expose comment les SIS de niveau étendu peuvent tirer profit des DP. Elle montre aussi des freins au niveau local et étendu impactant la qualité des SIS de niveau plus large.

4. Identification et analyse des plateformes de publication de DP en santé

L'évaluation de l'offre actuelle des DP en santé (Q2, §1) nécessite d'identifier des plateformes de diffusion de ces données. Celles-ci étant très disparates (volumes des données, mode de navigation...), les informations recueillies serviront essentiellement à dégager des tendances et des ordres de grandeur.

Parmi les nombreux catalogues qui recensent les plateformes mettant à disposition des DP, la *Semantic Web Company* les classe en cinq catégories : Autorités gouvernementales locales ou régionales ; initiatives locales ou régionales d'ordre privé ; autorités gouvernementales nationales ; initiatives privées nationales ; initiatives transnationales. La plateforme data.gouv.fr, lancée le 5 décembre 2011, n'est pas encore référencée !

Notre étude examine des initiatives transnationales et des plateformes pilotées par des autorités gouvernementales nationales.

4.1 Plateformes transnationales

Les plateformes transnationales sélectionnées appartiennent à l'OMS, autorité directrice et coordonnatrice des travaux liés à la santé, et à l'OCDE, qui appuie la mise en œuvre de politiques de qualité en santé et protection sociale.

Les statistiques fournies par l'OMS sont élaborées à partir de données de 194 pays, dont la France. Nous pouvons explorer plus de 25 bases de données, avec 80 700 jeux de données. Les plus récents (97 à ce jour) sont exportables aux formats CSV, XML, JSON, SDMX. Le Global Health Observatory comprend des séries statistiques complètes : il s'agit de 164 indicateurs couvrant 19 années (1990 à 2008), regroupés en 16 grandes catégories et exportables au format XLS.

La base de données de l'OCDE permet de comparer les systèmes de santé des pays membres. Les indicateurs couvrent l'état de santé, avec des variables de mortalité (espérance de vie, causes…), morbidité (maladies transmissibles, cancer…), et des déterminants non médicaux de santé (facteurs de risque, ressources, dépenses de santé…). Les 127 indicateurs, calculés sur 34 pays et 11 années, selon 26 unités de mesure (années, années d'écart, nombre de décès, nombre de cas...), sont exportables au format XLS. Les tableaux-clés plus récents sont exportables dans différents formats : HTML, PDF, XLS et XML.

4.2 Plateformes nationales

Les *plateformes gouvernementales nationales* 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. Trois plateformes sont considérées : data.gov (États-Unis, mai 2009), data.gov.uk (Royaume-Uni, janvier 2010) et data.gouv.fr (France, décembre 2011). Les volumes de données, toutes catégories confondues, n'étant pas comparables, trois sous-ensembles ont été retenus :

1. (US) : données du site data.gov, regroupées dans la Health Data Community
2. (UK) : données du site data.gov.uk, produites par le Department of Health
3. (FR) : données du site data.gouv.fr, répondant au mot clé « santé ».

Ces ensembles contiennent respectivement 249, 890 et 3442 jeux de données, dont 100%, 12% et 16% à couverture globale : pour (UK) les producteurs appartiennent au Health Department, mais les territoires couverts sont très hétérogènes ; pour (US), les données sont produites par les agences fédérales, tandis que les données (FR) sont produites par des institutions nationales et par des collectivités locales.

Concernant les *formats*, cette analyse confirme l'adoption majoritaire des formats XLS et CSV, XLS en (FR) et CSV en (US) et (UK). XML, présent en (US), l'est très faiblement en (FR), avec 2 « fils RSS ». RDF est absent de (FR).

Les *thématiques prioritaires* des ensembles (US) et (UK) focalisent sur les systèmes de soin et les aspects financiers (remboursements, coûts), avec une attention particulière à la transparence pour (UK). Celles de la plateforme (FR) sont très généralistes : Recensement de population, Catégorie socioprofessionnelle, Structure de l'emploi, Activité économique, Moyen de transport.

En effet, les thématiques sont corrélées avec des *questions d'organisation*.

Concernant *les producteurs*, le principal producteur de (FR) est l'INSEE (Institut national de la statistique et des études économiques). A la différence des autres plateformes nationales, le Ministère en charge de la santé fournit 15% des données de santé à couverture globale. En termes de *délais*, data.gouv.fr arrive plus de deux ans après les plateformes des USA et du Royaume-Uni et après que d'autres portails de DP se soient développés. Ainsi, en France les DP en santé sont fragmentées : le portail Assurance Maladie en Ligne diffuse de nombreuses données (consommation de soins,

prescriptions, activités et démographie médicale, données rétrospectives des dépenses) ; des organismes nationaux, en charge de la recherche médicale, du cancer, de la veille sanitaire ou de la prévention et de l'éducation pour la santé, contribuent au portail toutsurlenvironnement.fr (ministère en charge de l'environnement). De même, des organismes publics français alimentent les plateformes de l'OMS et de l'OCDE, mais pas encore la plateforme nationale.

5. Évaluation globale des apports actuels des DP à l'enrichissement des SIS

Malgré l'hétérogénéité des plateformes analysées, il est possible d'estimer globalement les apports actuels des DP à l'enrichissement des SIS (Q3, §1).

1) Pour quels types de SIS ? Avec quelles DP ? L'analyse ci-dessus conduit à estimer que les DP bénéficient surtout aux SIS de niveau étendu, national et international

– SIS décisionnels des classes 3 et 4 : DP de santé populationnelles, administratives et environnementales, largement géolocalisées, fournies par des institutions publiques de santé, nationales (comme le NHS) ou internationales (OMS), et par des organismes transversaux à divers niveaux (INSEE, OCDE)

– SIS opérationnels de la classe 2 : annuaires consolidés de professionnels et de services, données géolocalisées dans des thématiques connexes à la santé (notamment socio-économiques) et données géographiques, plus particulièrement sur les routes et les transports

– SIS de niveau local, de classe 1 : DP en santé pouvant faciliter des processus décisionnels, comme l'aide au diagnostic et à la prescription

2) Quels types d'enrichissement des SIS ? Les formats structurés de base, XLS et CSV, très largement adoptés, facilitent les études statistiques. La grande hétérogénéité des métadonnées associées aux différentes sources freine les approches d'interopérabilité basées sur les schémas. La faible représentation des formats riches sémantiquement (RDF notamment) rend l'ouverture des plateformes de diffusion de DP vers le Web sémantique loin d'être mûre, bien que souhaitée en santé publique.

3) Réutilisation des données publiques en santé : Quels freins ? Quelles tendances ? Les principaux *freins* à l'évolution des SIS grâce aux DP se situent, selon nous, au niveau organisationnel. Les quelques faiblesses liées à la « jeunesse » de la plateforme data.gouv.fr illustrent ce propos. Concernant les *tendances*, des citoyens et concepteurs d'applications en santé s'interrogent sur les convergences possibles des différentes plateformes de diffusion des DP de santé, les démarches, les délais, et, plus globalement, sur les mesures de ces évolutions. La convergence des plateformes publiques en santé apparaît particulièrement complexe de par les enjeux et la nature pluridisciplinaire du domaine.

6. Conclusions et perspectives

Les données publiques issues de plateformes de diffusion possèdent deux caractéristiques essentielles à l'évolution des Systèmes d'information : l'ouverture et la fiabilité. Cette étude exploratoire, réalisée à partir de plateformes institutionnelles nationales et transnationales, établit des corrélations entre les besoins d'évolution de différents types de SIS et ces données, des tendances sur les usages ainsi que des freins. Elle exhibe le fait que l'hétérogénéité est un problème majeur à résoudre pour assurer l'interopérabilité des données et des systèmes.

Nous envisageons d'évaluer plus finement ces plateformes, en développant des critères de comparaison plus pointus et une caractérisation plus précise des besoins de données pour les diverses catégories de SIS. La généralisation de cette étude à des SI non centrés sur la santé est aussi une perspective de notre recherche.

7. Références

Affset (2008). *Systèmes d'information en santé environnement. Enquête sur le croisement de données dans le champ santé environnement* http://goo.gl/ko8G9

Akoka J., Comyn-Wattiau I. (2009). *Vers l'ingénierie des évolutions,* dans *Ingénierie des Systèmes d'Information*, RSTI, série ISI, Vol. 14, n° 6/2009, pages 9-17

Boustany, J., Salzano, G. (2011) : *La réutilisation des données publiques : quels dispositifs ?*, 8e Colloque international de l'ISKO-France Lille 27-28 juin 2011

Chekli M., Lejeune A. (2009) *Systèmes d'information en santé : Classification, avènement du dossier santé personnel et apport de la science des services*, 14e colloque de l'AIM, Marrakech (Maroc) http://goo.gl/BwMvF

ENPC (2011) Ecole des Ponts ParisTech, *Les Données publiques au service de l'Innovation et de la Transparence. Pour une politique ambitieuse de réutilisation des données publiques*. http://goo.gl/6tf6z

Fieschi M.. (2009), *La gouvernance de l'interopérabilité sémantique est au cœur du développement des systèmes d'information en santé*, Rapport à la ministre de la Santé et des Sports

HCSP (2009), Les systèmes d'information pour la santé publique, décembre 2009 http://goo.gl/1i42j

Heath T., Bizer C. (2011), *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. http://goo.gl/enYQr

Leitzelman M. (2009), *Information Semantic Integration through Communities of Intelligence online*, http://goo.gl/yscG

Salzano, G. (2008), *Dimension géographique des Systèmes d'Information de santé*, in Nouvelles Organisations des Systèmes de Santé, SDM, Vol. 11, n° 3-4/2008, pages 153-168

Shadbolt, N. (2010), *From Data to Decisions: The Power of Information in the Age of the Web of Linked Data*. In: Royal Signals Institution Annual Seminar, HQS Wellington, London. http://eprints.ecs.soton.ac.uk/21624/

Wager, K.A., Lee, F.W., Glaser, J.P. (2009), *Health Care Information Systems: A Practical Approach for Health Care Management*, Jossey-Bass