
EPSI platform

European Public Sector Information Platform

Topic Report No. 2012 / 12

Open Data Standardization before publication?

Author: Katleen Janssen, Tom Kronenburg

Published: October 2012

Keywords

PSI, Public Sector Information, standardization, linked data, open data, file format, format

Abstract

In this topic report we explore the following question: Is it better to publish data 'as is' or to improve the data quality and to publish only after the data is highly standardized and usable within interoperable environments? On the one hand, we have the argument best formulated by Tim Berners Lee "Raw Data Now", but on the other hand many data holders prefer to hold the data until they are certain it's of sufficient quality. This Topic Report

discusses a number of ways in which data can be standardized and how these forms of standards benefit the re-use community.

Finally, we conclude that it's generally preferable to publish now. If the data has sufficient value, we observe that companies and civic groups will standardize the data themselves, and make it available to a wider public, without any cost to the PSB.

Table of Contents

Abstract

In this topic report we explore the following question: Is it better to publish data 'as is' or to improve the data quality and to publish only after the data is highly standardized and usable within interoperable environments? On the one hand, we have the argument best formulated by Tim Berners Lee "Raw Data Now", but on the other hand many data holders prefer to hold the data until they are certain it's of sufficient quality. This Topic Report discusses a number of ways in which data can be standardized and how these forms of standards benefit the re-use community.

Finally, we conclude that it's generally preferable to publish now. If the data has sufficient value, we observe that companies and civic groups will standardize the data themselves, and make it available to a wider public, without any cost to the PSB.

1 Introduction

Governments and public bodies hold an enormous amount of data that is of value to other public bodies, companies, organisations and the general public. Increasingly, these data are made available via national, regional or local portals, enabling them to be used for many different purposes, stimulating economic growth, transparency, participation and innovation.

When contemplating making their data available, public bodies are confronted with many different demands that cannot always be met, and they have to prioritise. One of the areas in which this becomes clear is the conflict between the demands for 'raw data now' and the requirement for public bodies to provide standardised data and ensure interoperability. This topic report takes a closer look at this dilemma and tries to make some recommendations on how to tackle the issues involved.

2 The question: raw data now or standardised data later?

'Raw data now' is one of the main credos that started the open data movement. It was first used by Tim Berners Lee in a talk at TED in 2009[1], who was in his turn inspired by a 2007 blog from the Open Knowledge Foundation's Rufus Pollock that asked 'Give us the data raw, and give it to us now'.[2] Applied to government data, this means that governments should make their data available in whatever format they exist, and with whatever inaccuracies, flaws, or gaps they contain. Hence, the data should be made available 'as is'.

On the other hand, increasing emphasis is being put by the open data community on the quality of the (open) data and the use of standardized data formats (preferably open

1 See http://www.ted.com/talks/view/lang/en//id/484.

2 See http://blog.okfn.org/2007/11/07/give-us-the-data-raw-and-give-it-to-us-now/

standards). It is indeed much easier to re-use data that are well-formatted, since they are easier read, easier understood by re-users and also easier manipulated and easier stored. In fact, in order to maximize the effect of opening up their data, it is often recommended that governments publish PSI in open formats, or even as Linked Data (a type of data in which not only content but even context is stored according to (open) standards.

However, it is also clear that many governments do not have their data organized according to perfect open standards. Often, datasets are stored within proprietary systems, stored according to proprietary standards. When opening such data, the data owner must make a decision whether to reformat the data, as well as deciding whether or not he wants to add to the general quality of the data (timeliness, completeness, etc). Hence, making government data available in a standardised form takes time, effort and money, and, as a consequence, conflicts with providing 'raw data now'.

It is clear that both demands are important, and in an ideal world, governments would collect and create their data already pre-formatted as linked data so that they can also be made available immediately to the broader public for re-use. However, currently this is not yet possible. So what priorities should the public bodies set?

2.1 The arguments for standardized data

PSB's only provide High Quality Products.

Many PSB's are in the business of delivering high quality government services to citizens and companies. They are focused on quality, and are aware that low-quality services can lead to problems for citizens, sometimes resulting in court-cases and bad publicity. Being risk-averse can be seen as a quality aspect of PSB operations.

When opening up data, these PSB's also want to deliver a high quality service. They want to know for sure that a dataset is complete, that is easy-to-use and understand, and that the data is updated as frequently as possible. This is a great benefit for re-users. Correct data, often updated and complete gives a very nice foundation to build services on. If a re-user is certain about the data source, he'll be much more likely to start re-using the data and build commercial products on it.

The ease-of-re-use

Another aspect of the quality of data, is the form in which it is delivered. For every programmer, it's very convenient to be able to immediately understand a file or datastream, and start reading it. If data is published in a .xls or .doc file, every programmer will understand that files like this will have to be read by Microsoft Office or products that also understand these (proprietary) standards. XML, HTML, CSV or plain .txt-files (given that they are also formatted according to the appropriate standards) can be read by a huge variety of programs, and can sometimes even be automatically processed by computer programs (this especially goes for Linked or Semantic Data).

In other words, well formatted data will have a much bigger chance of being re-used, because it is much more accessible to programmers. The format of a given dataset is even part of the OKF's "open definition" (http://opendefinition.org/okd/) in which it is said that

"The work must be provided in such a form that there are no technological obstacles to the performance of the above activities. This can be achieved by the provision of the work in an open data format, i.e. one whose specification is publicly and freely available and which places no restrictions monetary or otherwise upon its use."

Needing to buy licenses to use expensive proprietary data would definitely entail such monetary restrictions.

In effect, it is very smart for a government to publish data in a convenient open format, and they should ideally strive to publish data in a linked format!

Is time an important factor?

Many PSB's argue that the abovementioned quality aspects are of paramount importance for succesful PSI re-use. They argue, correctly, that it also takes time to organize these aspects. It also takes money, efforts and usually some organizational change. Unfortunately, this all results in a delay in delivering the data to the public. However, for these PSB's, time is usually a lesser problem.

2.2 The Argument for raw data

Raw Data is original data

The most important characteristic of 'raw' data, is that it is original. There has been no condensation or summarization of the data. Individual data points are recognizable and available for analysis. For many purposes of re-use, such detail is of paramount importance. Statistical data can be analysed in more ways if the re-user can re-use the original research data, video and audio can be more easily remixed when the re-user has the original tracks, energy usage data is more valuable when it's more finegrained. (i.e. house level is more valuable then street or even neighbourhood level). A re-user can never really know what 're-formatted' looked like originally. Raw data is therefore a more reliable source.

We'll unlock the value

Professional re-users and especially the high-tech community are not afraid to invest in reformatting PSI-data to fit their own purposes. If the value of the dataset is high enough, they will find a way to make it work. If the dataset is formatted to proprietary standards, published in closed formats they will find ways to reverse-engineer, scrape or otherwise collect the data.

Re-users argue that even when the data is delivered according to the most open of formats, they still need to adapt it for their own purposes. The cost of understanding the data and reformatting are often relatively low, compared to the possible value of the data.

Raw data can be delivered immediately

Most important however, is the fact that entrepreneurs, students, researchers and activists don't want to wait for data. They argue that the PSB's arguments for withholding easy access to the data, are stalling tactics used to protect the data as sources of income or protecting their information advantage. Raw data can be delivered with most ease. If a PSB doesn't want to deliver 'raw' data, there can be no other excuses, they just don't want it[3].

3 Whether or not there is a pretext can not be determined. Here we merely state an often heard

3 Background: what is open data about?

In order to be able to discuss the issue of standardised open data versus raw data, some more insight is needed in the concepts that are on the table: raw data, standards, open standards, open data, linked data, and interoperability.

3.1 Raw data

Raw data can be described as data from an original source, that have not been processed for further use. They are stored in a file or a database and can be processed manually or by a machine.[4] Raw data can sometimes be distinguished from value-added data or information, representing data to which structure, taxonomy, or context has been added. A distinction between data and information is that data can be automatically manipulated and processed by a machine, whereas information presupposes the involvement of a cognitive agent. Data is potential information. It becomes information from the moment it is understood by a certain person en has decreased its uncertainty or increased its knowledge.[5] The concept data does not have to be understood in relation to its receivers, contrary to information.

With regard to the open data debate, raw data is often also considered as data provided 'as is', without any quality guarantees, any 'cleaning' or standardisation, or an API to provide the data.

3.2 Standards

In very general terms, a standard can be described as an agreed, repeatable way of doing something[6] or a set of rules for ensuring quality.[7] Generally, it is laid down in a document established by consensus and approved by a particular standards organization. ISO defines a standard as "a document that provides requirements, specifications, guidelines or characteristics that can be used consistently to ensure that materials, products, processes and services are fit for their purpose".[8] According to European Directive 98/34/EC[9] a standard is:

"a technical specification approved by a recognised standardisation body for repeated or continuous application, with which compliance is not compulsory". Such a standard is either an international standard (adopted by an international standardisation organisation and

complaint.

4 See http://www.wisegeek.com/what-is-raw-data.htm

5 See P.B. HUGENHOLTZ, Auteursrecht op informatie, Deventer, Kluwer, 1989, 10.

6 See http://www.bsigroup.com/en/Standards-and-Publications/About-standards/What-is-a-standard/; http://www.cen.eu/cen/NTS/What/Pages/default.aspx.

7 See http://www.etsi.org/WebSite/Standards/WhatIsAStandard.aspx.

8 See http://www.iso.org/iso/home/standards.htm.

9 European Directive 98/34/EC laying down a procedure for the provision of information in the field of technical standards and regulations and of rules on Information Society Services.

made available to the public); a European standard (adopted by a European standardisation body and made available to the public), or a national standard (adopted by a national standardisation body and made available to the public).

Standards have many advantages: they can facilitate compatibility, coordination and communication, reduce complexity, bring advantages of large scale production, and increase production efficiency, but also reliability.[10] They also have some drawbacks: they take time and effort to develop and implement; they can be a threat to individuals' freedom to act and innovate; or they can encourage competition for supremacy between companies or even standardisation bodies.[11]

With regard to data, a data standard can be considered an established norm or requirement as to how a dataset is constructed. It generally includes criteria about the file format, the naming conventions, the quality of the data, the attributes that are included in the file or dataflow.

Related to data standards are data specifications and data models. Data specifications provide "a computer-readable description defining the data structure - enabling automated mechanisms for data management".[12] A data model is a "conceptual representation of the data structures that are required by a database", including the data objects, the associations between data objects and the rules which govern operations on the objects.[13]

3.3 Open standards

An Open Standard is a standard that can be used by everyone under the same terms. It is usually created by a large forum in which anyone can participate.

There are several definitions of open standards, some of which have been laid down in national or European legislation or policy documents. According to the 2004 European Interoperability Framework for pan-European e-Government services, minimum criteria to be considered an open standard are:

> the standard is adopted and maintained by a not-for-profit organisation and its ongoing development occurs on the basis of an open decision-making procedure available to all interested parties (consensus or majority decision etc.);
>
> The standard has been published and the standard specification document is available either freely or at a nominal charge. It must be permissible to all to copy, distribute and use it for no fee or at a nominal fee.
>
> The intellectual property - i.e. patents possibly present - of (parts of) the standard

10 See N. Brunsson et al. (2000). A world of standards. Oxford: Oxford University Press.

11 Ibid.

12 GS SOIL WP4, "D4.1 Theme specific test cases for developing "data specifications for spatial soil information, http://www.gssoil-portal.eu/Best_Practice/GS_SOIL_D4%201%20_theme%20specific%20test%20cases.pdf

13 http://www.liberty.edu/media/1414/[6330]ERDDataModeling.pdf.

is made irrevocably available on a royaltyfree basis.[14]

Interestingly enough, the new ISA Interoperability Framework, adopted in 2010[15], no longer uses the term open standards, but rather refers to formalised specifications and open specifications[16], which are also not defined.

While the term open standards primarily received a lot of attention a few years ago, it still remains on the political agenda today. For instance, the British Government is currently holding a consultation on open standards, based on the idea that on the one hand, the cost of the Government's IT is currently too high and needs to be reduced and there is a lack of diversity in government IT contracts; and on the other hand, it is difficult to transfer information and data across government boundaries and systems due to a lack of interoperability between products and services.[17]

3.4 Interoperability

Interoperability can be described as the ability of diverse systems and organisations to work together (inter-operate). Again, many definitions can be found in literature and policy documents. For instance, Decision 922/2009 of the European Commission on interoperability solutions for European public administrations (ISA)[18] defines interoperability as "the ability of disparate and diverse organisations to interact towards mutually beneficial and agreed common goals, involving the sharing of information and knowledge between the organisations, through the business processes they support, by means of the exchange of data between their respective ICT systems". The INSPIRE directive also holds a definition of interoperability, in the context of spatial data sets and services: "the possibility for spatial data sets to be combined, and for services to interact, without repetitive manual intervention, in such a way that the result is coherent and the added value of the data sets and services is enhanced".[19]

Several levels of interoperability can be envisaged. The European Interoperability Framework names four levels:

Legal: Aligned legislation so that exchanged data is accorded proper legal weight;

Organisational: Coordinated processes in which different organisations achieve a previously agreed and mutually beneficial goal

Semantic: Precise meaning of exchanged information which is preserved and understood by all parties; and

14 http://ec.europa.eu/idabc/servlets/Docd552.pdf?id=19529

15 Com(2010) 744 final

16 see http://jfopen.blogspot.be/2011/01/new-european-interoperability-framework.html

17 http://www.cabinetoffice.gov.uk/resource-library/open-standards-open-opportunities-flexibility-and-efficiency-government-it.

18 http://ec.europa.eu/isa/documents/isa_lexuriserv_en.pdf.

19 Directive 2007/2/EC

Technical: Planning of technical issues involved in linking computer systems and services.[20]

3.5 The Semantic Web

The Semantic Web provides a common framework, presently under development within the World Wide Web Consortium[21], that allows data to be shared and reused across application, enterprise, and community boundaries, by attaching sematic information to discrete datasets.[22] In this way, it should become possible to attach contextual meaning to data, facilitating its interlinking and interpretation.

The semantic web standards are also the basis for Linked (Open) Data.

4 Publishing standardised open data

The availability of standardised open data can be a big enabler for re-users. They can use widely available software for easy extraction, manipulation and back-up of the data. In addition, standards make it easier to understand both the semantics and the syntax of the data. Unfortunately, it's not always possible to unambiguously state whether a dataset is standardized or not. Standardization is more of a gliding scale, where the easiest form of standardization is adherence to a generic file standard, while perhaps standards that prescribe not only form, but also a specific type of content (of a specific quality, or with a specific understanding) are the most standardized. The higher a dataset ranks on this scale, the more interoperable it becomes.

Open data are made available under several types of standards, which can relate to different aspects of either the data or the files in which they are contained. The most general standards are file standards. For instance, data can be formatted according to .csv, .txt or perhaps even .xls or .odf definitions. These standards only apply to the syntax of the file. Even though many programs exist that can be used to open the files and manipulate the data that are contained within, the data themselves are not standardised. In addition, while some of these file standards may be considered open, this is not the case for all of them.

Some file standards are closely tied to the type of data that is being published. To a data owner, it will be immediately clear whether or not these standards are suitable to be used in or for a specific dataset. Examples include Google's KML format for geocodes, .MP3 and other audio files for sound data, etc. Even though the difference with the 'regular' file formats is somewhat academic, the main distinguishing factor with these file types is the presence of a 'context'. E.g. music data might be expressed (even though not easily played) as a spreadsheet, but a spreadsheet will not be stored as an MP3-file.

20 http://ec.europa.eu/isa/documents/isa_annex_ii_eif_en.pdf.

21 See http://www.w3.org/2001/sw/

22 http://en.wikipedia.org/wiki/Semantic_web#cite_note-W3C-SWA-1.

Other standards relate to the representation of the attributes and characteristics of the data. An example of such a standard are the Dublin Core standard (which can be used within the Resource Description Framework)[23]. They are used to add metadata to documents and webpages. Elements include title, creator, subject, description, publisher, contributor, date, type, format, identifier, source, language, relation, coverage and rights. However, Dublin Core does not provide any definitions for the different elements, nor does it standardise the way in which the elements are filled out. Some of them have schema's, while others do not provide any options and allow any free text. Another example of metadata standards can be found in INSPIRE[24], based on the ISO 19115 standard[25]. The INSPIRE metadata standards are freely available, while the ISO standards are only available for a fee.

The INSPIRE metadata standards not only define which type of elements need to be included, they also define how these elements should be completed, and provide a standard on how the data itself should be presented and even on the quality of the data. Such standards can also be referred to as data specifications.

Finally, there are standards about data that not only define the syntax, but also the semantics of the file. In Linked Data files we find the data that is being published, but also a reference to the relationship the data has with the 'web-of-things'. The file format for linked data files is RDF, often serialized into XML files. RDF[26] allows the publisher to link to the semantic web, thus supplying a context for the data. This is due to the fact that the RDF contains unique identifiers (uniform resource identifiers, i.e. URI's) to entities or relations that are stored within ontology-databases such as DBPedia. Linked Data are seen as the highest level of open data, because they enable true interoperability. It would no longer be necessary for a user to interpret the data, because the software could derive the meaning of the data from the context supplied by the URIs in the RDF (coupled with the other information contained in the ontologies in which the URIs are defined.). Data that is published as Linked data uses both the standardized ontologies available on the web (cf. DBPedia) as well as the metadata standards for describing the data contained in the file. More information on the metadata standards can be found in either the ISO 11179 standard, or the Dublin Core set of best-practices (http://dublincore.org/metadata-basics/)

5 Working with non-standardized open data

It is also possible to publish non-standardized open data. As we have already stated above, most government data is actually published in a standardised file-format, but without any further levels of standardisation. Fortunately, such non-standardised data can still be re-used and 'cleaned up' in various ways.

Government data is often published in the form of excel files. While such files are generally

23 http://dublincore.org/.

24 See http://inspire.jrc.ec.europa.eu/index.cfm/pageid/101.

25 http://www.iso.org/iso/catalogue_detail.htm?csnumber=26020.

26 http://www.w3.org/RDF/

accessible for any user, the lay-out and presentation is usually not standardised. For instance, columns may or may not have headers indicating their content; cells may be merged making it difficult to select multiple cells; names or words may be misspelled, making it difficult to query the file, etc.

Manual analysis and cleaning of the spreadsheet tables or databases is often required to make the data usable. This is also explained in our topic reports on data journalism and budget data[27]. For instance, the OpenSpending project spends most of its time on manually rearranging data stemming from many countries, regional and local authorities, all provided in many different formats. While manual analysis may work for small data sets, it is difficult to use for large data sets of for data sets that are frequently updated. The cost of performing repetitive manual analyses would prevent the development of a feasible business model.

Some examples exist where the publication of excel-sheets has led to a valuable and viable re-use of the data contained in these files. For instance, the Dutch company 10000scholen.nl, which runs a website with PSI-based information on the Dutch public elementary and secondary schools, has created an automated process to load such excel data into their system. They can do this because the PSI-holder, DUO (Dienst Uitvoering Onderwijs), has promised to keep publishing its excel files in such a way that the meaning of a specific cell or column never changes. If the data holder changes that policy, 10000scholen.nl will have to change their import procedure, at considerable cost.

Compared to PSI incorporated in excel-sheets, using PSI stored as a text document is much harder to use. To extract data or information from such text documents, the Linked Data community has started to develop a number of parsing engines. These are pieces of software that will 'read' the text, will extract every piece of information it contains and store this information in a database. The software will use mainly two different ways to extract facts from the text:

The software is able to recognize subject, object, nouns, verbs and other linguistic constructs. It will 'understand' that words like "is", "are", "more than", "less than", "in" etc, have a specific meaning and will be able to derive a relationship between two concepts contained in a sentence. Consider the sentence: "Brussels is the capital of Belgium." Most parsing engines will derive that [BRUSSELS] is [CAPITAL] of [BELGIUM]. The software does not need to know whether Brussels is a city, a butcher or a car, it just needs to record that the relationship between Brussels and Belgium is "Capital of". Later, the software will be able to query the database for [Concepts] that are [Capital of] [Belgium], to which the software will find that Brussels is such a Concept. The database might also produce a list of [concepts] that are [Capital of] [concept]. Given that the parser has read more text than just the one line, it might come up with a list of capitals it has read about.

In combination with this language technology, the parser might be connected to a database with established facts. Such a database could be DBPedia or any other linked data - database. DBPedia will know that Brussels is, in fact, a city. If the parser would encounter a

27 http://epsiplatform.eu/analysis

word like Brussels (or any other noun or verb), it would try to 'look up' this word in the DBPedia, and find that the concept Brussels is a city, and is not only the capital of Belgium, but also the capital of Europe, the capital of the "Brussels-Capital region" etc. Given that the parser is sufficiently certain about the meaning of a word (from the context), the parser can use this information to further understand the text. It will now also know that Brussels is e.g. not a butcher or a car.

This parser technology is used by a variety of companies, and is used in European projects like LOD2. It is important, because in time, it will allow automated processes to be built on top of unstructured datasets. However, the technology is not yet commonly applied, especially in relation to PSI.

In a future report on Linked Open Data, we will delve deeper into this semantic (parser) technology, but for now it is enough to understand that there are ways to automatically 'parse' non-structured, non-standardized texts into linked data, formatted according to RDF standards and using Linked Data ontologies (standardized meaning) and/or URI's for identification.

5.1 Citizen activity - the further standardization of open data.

We have often seen that high value data has been published in a less then ideal format for re-use. The emphasis within the PSB publishing the data is on openness, not on re-usability. However, provided that the data is sufficiently valuable, we have also seen that a number of companies and civic groups have started to reformat the data and building datasets or datastreams that are of higher quality.

One such example is the Hungarian (now EU) project Parltrack. Parltrack scrapes data from a number of European Parliament (EP) websites, and republishes that information in a very re-usable way. The file standards have been improved (PDF and HTML to XML/RDF), and Parltrack has implemented a standardized method of recognizing an EP-dossier. This turns the EP-data essentially into a form of Linked Data.

Another example comes from the UK, where the OpenCorporates project has published a database of linked Corporate data that is free to re-use, and which brings together the Chamber of Commerce data from a number of countries both inside and outside the EU. The project team has scraped pdf's, copied databases and transformed datasets in order to create one big resource that (in time) aspires to contain 'every company in the world'.

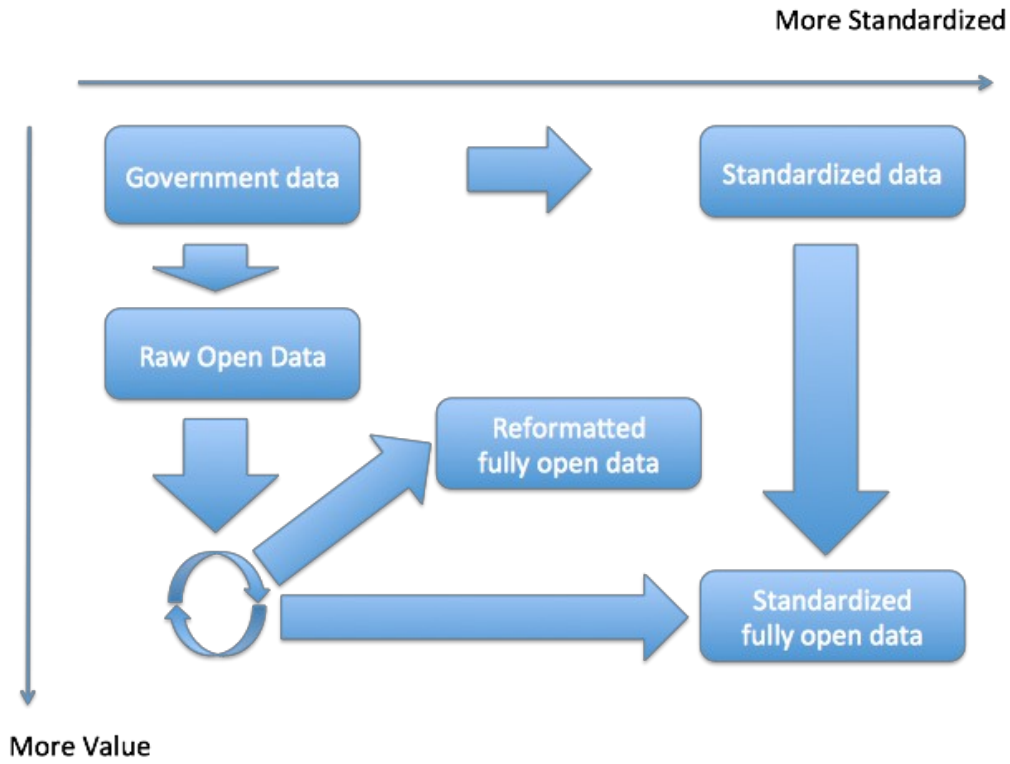
From the Netherlands come the examples of 4 separate companies that add value to car data provided by the RDW (the official Dutch car register). They make a profit out of combining PSI with commercial datasets and reformatting car data in order to be more easily brokered by insurance companies.

The three examples presented here show that as-is data, sometimes not even published as datasets (but rather as scrapable webpages) can be used to create more standardized datasets,sometimes available for free, sometimes for profit. However, it also shows that if the data has sufficient value, others will come and improve the data. The PSBs publishing the data don't necessarily have to concern themselves with improving the data, and

standardizing it, before releasing it for re-use.

6 Conclusions: which path to choose?

We have sufficiently established that publication of open data by data holders comes in many shapes and sizes. The discussion on whether these PSBs should publish first and improve later, or standardize and improve first and publish later is however, not yet answered in this paper. Conceptually, the problem can be described by the diagram below.



The diagram shows the different paths that can be taken to make government data (PSI) available to the public. The PSB that wants to publish its data, can follow one of two options. Either it can publish the data as raw data, or it can make efforts to standardise the data before it is published. Both approaches have their advantages and drawbacks.

The advantage of public bodies publishing the data 'as is', as raw data without any standardisation or quality improvement, is of course that the data is immediately available to whoever wants to make use of it. However, the data has been created for a particular purpose within the public body and is in most cases conceptualised as data that will only be used internally, without any consideration of its potential use for other parties, either within or outside the public sector. This means that the data may not immediately be useful or of sufficient quality for the purposes of other potential users. The group of re-users might be limited to those with deeper understanding of (processes within) the PSB publishing the data, or those with specific technical skills. However, we have seen many projects in which these technically adept select few have transformed the data into a much better, much more interoperable, more standardized dataset. The public, in those cases, have greatly advanced the re-usability of the PSI.

Conversely, if the data is 'cleaned up' by the publishing PSB, its quality is improved and/or it is standardised before it is published, then its potential for immediate utilisation and valorisation will improve considerably. However, standardising the data before publishing it, takes a lot of time and effort from the public bodies. For instance, while the INSPIRE programme will lead to a wealth of data becoming available, the process will have taken around 15 years by the time it will be fully completed. Moreover, the focus on standardisation, data specifications, services and technical requirements in general has caused the impression among some that the actual focus of INSPIRE - data sharing - has been lost.

Some public bodies consider their data not of sufficient quality to make it available to other parties, and worry about their own reputation or liability. Hence, they want to make sure the data is of good quality before they publish it. This also includes standardisation efforts. However, this entails the risk that public bodies may start using the need for standardised data as an excuse to postpone the publication of the data, while it could already be very useful in its non-standardised, non 'quality-assured' form. The consequence of this policy choice would be that the full economic, civic and innovation capital of PSI re-use is significantly delayed.

Ideally, public bodies would take into account the possibility of further dissemination towards third parties from the moment they begin creating or collecting data. In this way, data could immediately be shared in a standardised form, or as linked open data or any other highly re-useable format. This would require the concepts behind open data to be adopted throughout the entire data lifecycle, and especially be present within IT purchase processes.

However, as this is currently not the case in many public bodies, it is more beneficial for the citizens, the private sector, civil society, and the public bodies themselves that the data is made available 'as is', allowing others to play a role in cleaning up the data and making it re-usable. This community can then in its turn also 'give back' and play a role in the standardisation of the open data. This allows for quicker valorisation of the value that lies within PSI, more possibilities for re-use and a generally better European society.

About the Authors

Katleen Janssen (1978) is a postdoctoral researcher in information law at the Interdisciplinary Centre for Law and ICT at the Faculty of Law at the KU Leuven and a professional support lawyer at time.lex law firm. Katleen specialises in access to and use of Public Sector Information, open government data, and SDI- and GIS-matters. This includes policies promoting the availability of information and policies restricting such availability, e.g. privacy protection, intellectual property rights, etc. In 2009, Katleen obtained her Phd with a thesis about the legal framework for the availability of public sector spatial data, mainly dealing with the relationship between INSPIRE, PSI and access to environmental information. For more information, see http://www.law.kuleuven.be/icri/people.php.

Tom Kronenburg is a consultant with Zenc B.V. based in the Netherlands. He specialises in information as a solution to societal problems. Tom is one of the curators of the EPSI Platform website and travels throughout the European Union to connect PSI holders and re-users, citizens and governments.

Copyright information