Recommendations for formats for PSI re-use

François Bancilhon

ISM and Data Publica

GFII ("Groupement Français des Industriels de l'Information") is a group representing the key industry players in the information sector in France. It covers a broad spectrum of verticals (legal, economic, scientific, business, etc.). Within GFII, a committee is in charge of the PSI issue. This committee includes, together with industry representatives, public organizations representing some of the key publishers of public data (for instance DILA for legal and administrative, IGN for geographical ant INPI, for IP and companies information, etc.). This committee recently issued a set of recommendations concerning formats to be used for PSI publication. A French version of the document can be found at http://www.gfii.asso.fr/fr/document/reutilisation-des-informations-publiques-et-formats.

This position paper presents most of these recommendations and resets some of them in a European setting. Because of choice and extensions of the recommendations I made, it should not be considered as an official position of the GFII, but the GFII input should certainly be acknowledged.

1Physical and conceptual levels in PSI

We consider that information can be looked at two different levels

1. physical allowing essentially presentation and display of documents
2. and conceptual, allowing for the understanding and analysis of the semantics of the documents and data

The physical level is generic and does not depend from the domain addressed by the document. Physical formats for text include for instance ascii, xls, csv, xml, and html, etc, for image, jpeg, and gif , and for video, mpeg-4.

The second level (conceptual) is in general dependent of the specific domain addressed by the document (eg geographic, transportation, legal, economic, etc.). It is in general based on a data model taking into account the addressed vertical.

- For instance, in the case of geographic, the INSPIRE Directive yielded a conceptual model, whose usage is to become mandatory for the exchange of data within the EU. INSPIRE defines 32 themes, each one of them with a specific model.
- A similar approach was used for transportation with the ITS initiative (Information and Transport Systems) http://ec.europa.eu/transport/its/road/action_plan/action_plan_en.htm
- For financial data, the same approach is used in XBRL
- Finally for the cultural domain, the OAI protocol is commonly used for exchanges between libraries, archives and museums www.openarchives.org

2- Recommandations

1-In support of open formats

PSI should be published both in an open format and in a format compatible with the most common available tools on the market.

Open format is a good and desirable thing. Requesting that it is only available in that format would make it not usable for a large class of people. For instance, while odt is a great format, and requesting that documents are published using the odt format is a good idea, publishing it only in odt would make it unusable to a large class of users.

2-Meta data standard by verticals

We recommended that meta data standards be defined and enforced by vertical sector to allow for the proper re-use of data. Key vertical sectors such as legal, economic, etc. should be covered. These standards should be established by standard bodies including representatives of academia, industry and public bodies. These standards should be based on existing recognized standards such as the Dublin Core. They should clearly be done at the European level. Once again, the INSPIRE directive is an example of such an approach.

3-Anonymizing data

Anonymisation is an important issue. In some countries, to respect privacy, PSI containing the name of individuals can be published only after it has been anonymized. We consider important that the same rules and regulation apply to all across Europe. We consider further that it is important to define clearly the criteria that the anonymization process must follow and to put in place a certification of the anonymization process.

4-Providing source data with PDF data

PDF is an excellent format for data presentation, for the general public. It is however a very bad format for re-use, because it forces the re-user to reverse engineer the data., which is a counter productive exercise. Furthermore, in many cases the data in text PDF was produced from a source format which would be more re-usable (word, excel, powerpoint, etc.). So requesting to give access to the source of the document makes sense and would represent no extra cost for the publisher

Thus we recommend that the source file of the pdf, when it is available, be published at the same time as the pdf file

5-Privileging XML

Adopting XML as a publication format is a positive step from most PSI publisher. This format shouldthus be privileged. This adoption must be accompanied with a proper strategy in terms of the XML format (see the recommendation on meta-data standards).

6-Definitions of exchange formats

Many public bodies publish data using their internal format. This format is not focused on publication but on internal use of the data. We recommend instead to define an exchange format that would be focused on data re-use. These formats could be jointly elaborated by the public bodies and the re-users. Many such exchange formats have already been defined by XML DTDs in various verticals. They have the benefits of providing a good tool for insuring interoperability between heterogeneous systems.