CIDOC2014, Dresden, 8-11 September 2014

Quality for Linked Data. What is it and how can it be made?

Jef Malliet

Erfgoedplus.be, PCCE, Provincie Limburg
Universiteitslaan 1
B-3500 Hasselt
Belgium
Tel: +32 11 238 384
E-mail: jef.malliet@limburg.be
http://www.erfgoedplus.be

Keywords: linked data, semantic web, data quality

Summary

Everybody agrees that in order to be useful in a Linked Data environment, 'metadata' for cultural heritage objects must fit some quality requirements. Although quality of information is a subject of research since several decades, it is as yet unclear what it means for the semantic web. Furthermore, cultural heritage metadata is normally conceived in the first place for collection management purposes and not for sharing or publication, let alone for linking. There is much confusion about where in the workflow the required quality can or must be made and by who. This paper provides some clues about the elements of data quality for LD and how it can be made, based on experience with heritage Linked Data initiatives: the local Erfgoedplus.be and the broad Europeana.eu.

Dimensions of quality

Almost as long as there have been information systems, there have been studies about the quality of information. Usually these describe quality along a set of dimensions, the most recurrent seem to be:

- Accuracy ([1 3 4])
- Completeness ([1 2 3 4 5])
- Consistency ([3 4 5])
- Correctness ([1 2 4 5])
- Currency / timeliness ([1 3 5])
- Integrity ([5])
- Relevance / meaningfulness ([1 2])
- Safety / reliability ([2 3])
- Understandability ([3])

Input for this list comes from several publications[1] [2] [3] [4] [5]. It is definitely not exhaustive, as is also not the list of relevant publications (the intention is merely indicative). Each of these sources typically makes a list of 4 to 6 dimensions. Clearly the selection of relevant criteria and the way they are interpreted should depend on the contexts of origin and of usage of the information. However, such context is often not readily identified in the texts.

Four stages of re-use

Digital heritage information in collections has several uses, or levels of usage: collection management, publication, exchange, linked data. Typically these can be seen also as levels or stages, often even in chronological order, of awareness about the usefulness of heritage information. Indeed, each of these steps represents a significant enhancement of the way the information can be used. This in itself is an enrichment, but each step also adds new expectations and standards to be applied. The requirements of information quality will be different in each situation and gradually become more demanding for higher levels. One way to cope with these differences can be to produce separate datasets for each level. The information quality can then be made to match each level's requirements. But this is not very cost-effective. Usually the information is produced primarily for internal collection and information management purposes (the first level) and then the overall aim is to recycle or convert it, eventually enhance it, to be used at the higher levels.

[1] Richard Y. Wang, Henry B. Kon, Stuart E. Madnick, *Data Quality Requirements Analysis and Modelling*, Ninth International Conference of Data Engineering, Vienna, 1993, accessible at http://web.mit.edu/tdqm/www/tdqmpub/IEEEDEApr93.pdf

[2] Kuan-Tsae Huang, Yang W. Lee, Richard Y. Wang*, Quality Information and Knowledge*, 1999

[3] Karel Dejaeger, Jessica Ruelens, Tony Van Gestel, Joachim Jacobs, Bart Baesens, Jonas Poelmans en Bart Hamers, *Evaluatie en verbetering van de datakwaliteit*, in Informatie, November 2009, p. 8-15, accessible at http://www.researchgate.net/publication/241763540_Evaluatie_en_verbetering_van_de_datakwaliteit

[4] Maurice van Keulen, *Onzekere databases*, in Database Magazine, June 2010, p. 22-27, accessible at http://eprints.eemcs.utwente.nl/18030/01/DBM201004_Thema_Van_Keulen_117343.pdf

[5] Laura Sebastian-Coleman, *Measuring Data Quality for Ongoing Improvement: A Data Quality Assessment Framework*, 2013

Stage 1: Collection management

Digital heritage information, in particular in museums, is made for the purpose of collection management. Cultural heritage carries a public responsibility and there is an obligation to maintain an inventory of managed heritage objects. A good inventory must allow to identify the objects for administrative and security purposes. This basic functionality has been expanded, to include all descriptive and biographical information available about the object. It further developed into the repository where all documents, reports, photographs, etc. relative to the objects are gathered and linked. A good inventory is therefore now a full-fledged collection management system, a backbone for the processes for managing the collection (see SPECTRUM[6]), as well as a content management system, that allows to keep all relevant information organized per object in the collection, and a digital assets management system, including also scans and digital photographs.

Stage 2: Publication

The following step in the life of a digital inventory is marked by the possibility to make the collected information available for museum visitors, experts or general public. This could in the first place be done in house, to assist visitors with their visit, or by extracting relevant information to produce printed catalogues. The Internet brought interesting new ways of sharing such information, allowing cross-referencing with other objects through direct database searching or html. Typical for this stage is the development of suitable keyword lists or thesauri, to facilitate finding information of interest.

Stage 3: Exchange, aggregation

Proliferation of databases published on websites of each individual collection implied that users wanting to find information independent of the collection itself had to locate relevant websites, where they then had to get acquainted with the local search tools. Web search engines can help with the finding, but often produce many irrelevant results as well, and they cannot help with unifying the content. Hence the emergence of exchange services like portals, federated search and aggregators, specifically designed for heritage information. These required new information representation and exchange formats such as XML, Dublin Core, LIDO, but also ways of interpreting information content-wise across the collection boundaries. The need for unified thesauri is strong.

Stage 4: Linked Data

Yet one step further comes the desire to be able to link an object and the elements in its description to objects or concepts that are described elsewhere. This turns the information from a database with words into a network of interlinked concepts. Much effort is currently invested in the exploration and implementation of such a 'semantic' web. This requires unique persistent identifiers of objects (physical and digital), ontologies, concept-based thesauri, etc. The most useful reference objects would be places (geolocation), persons (or agents, incl. institutions and groups), events, besides the more general concepts which can be found in thesauri. Relevant technical developments include CIDOC-CRM, RDF, SKOS, OWL.

[6] Developed and maintained by Collections Trust, UK, 'SPECTRUM is an open and freely available collections management standard. It is recognised nationally and internationally, as the primary specification for collections management activity in museums': see http://www.collectionstrust.co.uk/spectrum

Quality for Linked Data

A helpful approach may be found in the Wikipedia article about the 'semantic web'[7]. The article describes the 'challenges' for the Semantic Web to include *vastness*, *vagueness*, *uncertainty*, *inconsistency*, and *deceit*. It further states that *'This list of challenges is illustrative rather than exhaustive, and it focuses on the challenges to the "unifying logic" and "proof" layers of the Semantic Web.'*

Note that the explanatory notes in the Wikipedia article seem to originate from a few specific environments, i.e. in particular medical. Furthermore, the possible solutions presented in general concern technical solutions, on the end-user side. Here it will be argued that technical solutions alone cannot solve the problems effectively, and that the information production side must definitely be made conscious of their role and influence in providing quality information. The concepts shall be reinterpreted within the cultural heritage environment. The Wikipedia texts are added under each dimension as a general introduction.

Challenge 1: Vastness

> *Vastness: The World Wide Web contains many billions of pages. The SNOMED CT medical terminology ontology alone contains 370,000 class names, and existing technology has not yet been able to eliminate all semantically duplicated terms. Any automated reasoning system will have to deal with truly huge inputs. [Wikipedia http://en.wikipedia.org/wiki/Semantic_web]*

Problems:

Besides a normal problem, due simply to existence (and rapid growth) of the amount of data, there are specific concerns related to duplicate information and selection or relevance of information.

Duplicate information can easily appear because

- proper unique identification of each object is not solved,
- there may be multiple sources of information about the same objects, lacking proper linkage,
- there may be multiple channels of publication of the same information.

Sometimes selection is proposed to reduce the risk of information overload. Only the most reliable and relevant information should be published. With more traditional, expensive forms of publication (such as books and magazines) there is a kind of 'natural' selection: only the strongest contributions survive. However, in the www it is much easier to divulge less relevant or less controlled information.

The Internet technology of the web 2.0 allows users to publish their own contributions or to add their comments to information published by others. Professionals fear that this may contaminate their scientifically more correct information. In general, web 2.0 adepts claim that this would regulate itself, because the entire user community can see to it that wrong or bad information is removed or corrected (e.g. Wikipedia). However, it must be noted that the

[7] http://en.wikipedia.org/wiki/Semantic_web, as accessed on 17 July 2014

user community around cultural heritage is not very large and then this auto-regulation will not work very efficiently.

Some claim that there is more control required. However, control would mean censorship, and this is against the rules of the web 2.0, and even against universal human rights, which feature the basic principles of open access to information, freedom of expression and respect for the rights of others.

Answers:

Duplication must be kept under control. Persistent identifiers are needed, and they must be assigned as closely to the source (the collection holders) as possible. These primary sources would need to be actively interested in following up on the derived products (aggregators and re-users) so that wrong re-use can be spotted and corrected. Care should be taken that information via different channels be communicated in a uniform and consistent manner.

Though user-generated content is not always trusted by the ‘professionals’, they usually appreciate that it can be useful and enriching. Other users may have valuable additions, even the non-professional volunteers who often can have very good knowledge about very specific subjects or items. The professionals want to be able to control and select the good parts. Also for this purpose, the collection holders as primary sources have a responsibility to follow up on the additions to their information as made by users, and eventually confirm or correct them.

Challenge 2: Vagueness

Vagueness: These are imprecise concepts like "young" or "tall". This arises from the vagueness of user queries, of concepts represented by content providers, of matching query terms to provider terms and of trying to combine different knowledge bases with overlapping but subtly different concepts. Fuzzy logic is the most common technique for dealing with vagueness. [Wikipedia http://en.wikipedia.org/wiki/Semantic_web]

Problems:

Information posted on the web has in general been produced within a specific context, for specific user groups. When it is prepared for aggregation or for Linked Data, the potential user base broadens and this context disappears. The semantics become uncertain and confused. Often it is even not clear to what object the posted information pertains: an original physical object, or a derived object at any stage, e.g. photographs, scans, transcriptions.

The information can be prepared with the help of authority files or thesauri. However these often have been composed within the same specific context, and rarely contain sufficient information to identify or clarify this context when the information is shared outside the source environment.

Users are not generally aware or knowledgeable about the context of the origin of the information they are looking at. They have varying expectations and may make alternate, incorrect interpretations. When they re-use such information, the result may turn less reliable.

In the source environment, certain background information may seem very obvious and be omitted from the recorded information. This results in the published information not being complete when seen independently by users that do not belong to the environment.

Answers:

Thesauri should be used that are broadly used and understandable outside specific contexts. The semantics, explained in 'scope notes', are very important in the thesaurus and should clearly define the concepts and their boundaries. When using domain specific authority lists, these should be able to refer/link to other thesauri of wider use.

Enrichment through thesaurus connections are best made at the source, or as close to the source as possible. The further away from the source, the less accurate the enrichment will be. At larger distance, interpretation is less certain and becomes more vague. Aggregators who try to provide enrichment by applying linking to external resources (usually in a machine-controlled manner) may be able to make good connections, but will inevitably also make many wrong links, reducing the overall reliability of the information.

When recording information about their heritage objects, collection managers should be aware of the usage of such information outside of the context of their own collection. When existing thesauri are used, specific care must be taken to use the concepts in a consistent manner, or when creating a thesaurus, it should carry a proper description of the concepts, also understandable outside the context of the collection. Implicit reference to contextual information should be recognized and made explicit when possible.

Challenge 3: Uncertainty

Uncertainty: These are precise concepts with uncertain values. For example, a patient might present a set of symptoms which correspond to a number of different distinct diagnoses each with a different probability. Probabilistic reasoning techniques are generally employed to address uncertainty. [Wikipedia http://en.wikipedia.org/wiki/Semantic_web]

Problems:

Facts about heritage objects are not always clear-cut. While titles, names of authors, publication data can be read directly from the documents themselves in libraries, this is not always true for artefacts or objects held in museums or archives. Much of the information about such objects is the fruit of study and interpretation. Two (or more) versions of similar information about the same object, as produced by different scholars, or by the same person at different moments can present diverging versions of the 'truth'.

For characteristics that are expected to be expressed in precise terms, such as time and location (e.g. place and date of creation of an object), such interpretation leads to estimations between more or less vague margins. The precision of observations and measured properties, or the measurement method applied are not always included with the values, which could as well be rough estimations or guesses. Particularly when the information is passed on to a secondary source (such as an aggregator), any specifications regarding precision can be lost in the conversion to another data format. Often the descriptions also reflect opinions, the source of which is not always indicated. All of these factors of uncertainty can be mixed.

Answers:

Uncertainty can definitely not be ruled out. The ‘truth’ has many faces and requires interpretation, backed by the author’s own history and experience. Ideally the end-user needs an indication of the degree of certainty of the information, or the trustworthiness of the source. The source of the information is therefore important metadata that should accompany any descriptive information which is not directly derived from the (digital) object itself. It should as much as possible be passed on each time the information is shared or re-used. The end-user should be given enough elements to be able to verify and judge how reliable the received information is. Any context information about the background of opinions or the methods of observation are very important to help evaluate uncertainties. Indicating diverging opinions in a description can enhance the appreciation of a source as trustworthy.

Challenge 4: Inconsistency

Inconsistency: These are logical contradictions which will inevitably arise during the development of large ontologies, and when ontologies from separate sources are combined. Deductive reasoning fails catastrophically when faced with inconsistency, because "anything follows from a contradiction". Defeasible reasoning and paraconsistent reasoning are two techniques which can be employed to deal with inconsistency. [Wikipedia http://en.wikipedia.org/wiki/Semantic_web]

Problems:

Inconsistency may appear in the content, due to the fact that information from various sources is merged into one product. Facts can be contradictory, subject to many interpretations, or interpretations changing over time.

It can also be generated by the way the information is registered. If normalization is not well taken care off in a database, information can be entered in the wrong fields, or mixed information or of different types can appear in a single field, thus making consistent transformation for re-use impossible.

The differing contexts of the sources can generate inconsistencies as well. Different contexts may handle other words, or other meanings, or another model of reality entirely. Likely there may be inconsistencies in the model itself, due to the fact that any ontology usually has to combine more detailed models from various viewpoints.

Answers:

Indication of the source of information should allow the end-user to clearly differentiate between diverging versions of information. This must help to evaluate the origin of the differences and give clues for construction of a proper opinion.

Standards are fundamental to help with achieving consistency in formulation and interpretation of information. Relevant standards include reference models, data formats and authority files, as well as proper guidelines for interpretation in specific contexts. These standards must then be followed seriously and as much as possible without adaptations to local situations.

Application of broadly accepted thesauri or common ontological models helps rule out improper interpretation or miscommunication between source and target users.

Challenge 5: Deceit

Deceit: This is when the producer of the information is intentionally misleading the consumer of the information. Cryptography techniques are currently utilized to alleviate this threat. [Wikipedia http://en.wikipedia.org/wiki/Semantic_web]

<u>Problems</u>:

The Wikipedia article seems to consider only bad intentions under this point. Authenticity of digital information is indeed hard to assess and demonstrate, but information can simply be wrong or unbalanced for various reasons and thus deceive the reader, whether intentionally or accidentally.

Information can deceive because it is old, and new insights have overtaken previous interpretations. The older versions can still co-exist on the web, on their own or in various derived forms.

Certain aspects can appear overemphasized, in particular when information from different sources is combined, thus distorting the overall perception.

Merged information from various sources can have mixed precision and quality, creating quality expectations on the reader's side that are not matched by all pieces of the information.

<u>Answers</u>:

As most of the problems of deceit emerge from the publication or aggregation actions, they can only be countered by good monitoring after publication or aggregation. It is mainly the contributors of the information who are best placed for taking corrective action. The views and feedback from end-users can definitely help to detect eventual problems. Feedback and interaction with end-users must therefore be considered seriously and used for improving the presentation.

To avoid deceit, the original information may need to be more appropriately interpreted and completed. This must be done by the original source of the information or as closely to the source as possible, in order to be reliable.

The source of information must be clearly indicated, so that the end-user can value the differences in quality and provide eventual feedback.

Precision of information should be indicated so that the end-user can recognize varying levels of precision.

Conclusion

This paper needs a conclusion, but it is certainly not a conclusion of the discussed issues. Throughout the discussion, there are a few points that keep returning, and thus merit special attention.

Quality of data is often approached as a technical problem that can be solved with technical solutions. However, the majority of issues discussed show that quality is prevalently a content issue, which cannot be generated in a reliable manner by automated information systems. Human action and interaction is fundamental for most quality issues.

Quality is made prevalently by or in collaboration with the authors of the information itself, usually in the environment of the collection holder. It is hard for others to add quality to existing information. Reference to the source of information is definitely an added value because it allows end-users to estimate the reliability.

Proper contextualisation is an important aspect of data quality. This implies indication of the sources of information as well as expressing the semantics in appropriate ways, avoiding context assumptions as much as possible. Semantic links (the links of Linked Data) should be made by who is familiar with the original context of the data.

Help from the end-user, as user-generated content or feedback, is much appreciated. However, as primary authority the collection holder should take the task to follow up and keep watch over the relevance and quality of such contributions.