JeDEM 4(2): 222-244, 2012
ISSN 2075-9517
http://www.jedem.org

The potential of metadata for linked open data and its value for users and publishers

Anneke Zuiderwijk, Keith Jeffery, Marijn Janssen

Anneke Zuiderwijk, Delft University of Technology, Faculty of Technology, Policy and Management, Jaffalaan 5, 2628 BX, Delft, The Netherlands, a.m.g.zuiderwijk-vaneijk@tudelft.nl, +31 15 278 6471

Keith Jeffery, Science and Technology Facilities Council, Rutherford Appleton Laboratory, Didcot, OX11 0QX, Harwell Oxford, United Kingdom, keith.jeffery@stfc.ac.uk, +44 1235 44 6103

Marijn Janssen, Delft University of Technology, Faculty of Technology, Policy and Management, Jaffalaan 5, 2628 BX, Delft, The Netherlands, m.f.w.h.a.janssen@tudelft.nl, +31 15 278 1140

Abstract: *Public and private organizations increasingly release their data to gain benefits such as transparency and economic growth. The use of these open data can be supported and stimulated by providing considerable metadata (data about the data), including discovery, contextual and detailed metadata. In this paper we argue that metadata are key enablers for the effective use of Linked Open Data (LOD). We illustrate the potential of metadata by 1) presenting an overview of advantages and disadvantages of metadata derived from literature, 2) presenting metadata requirements for LOD architectures derived from literature, workshops and a questionnaire, 3) describing a LOD metadata architecture that meets the requirements and 4) showing examples of the application of this architecture in the ENGAGE project. The paper shows that using metadata with the appropriate metadata architecture can yield considerable benefits for LOD publication and use, including improving find ability, accessibility, storing, preservation, analysing, comparing, reproducing, finding inconsistencies, correct interpretation, visualizing, linking data, assessing and ranking the quality of data and avoiding unnecessary duplication of data. The Common European Research Information Format (CERIF) can be used to build the metadata architecture and achieve the advantages.*

Keywords: Metadata, linked open data, LOD, open data, metadata architecture, requirements, elements, CERIF

Acknowledgement: This paper is related to the ENGAGE FP7 Infrastructure Project (An Infrastructure for Open, Linked Governmental Data Provision Towards Research Communities and Citizens), that started in June 2011. The main goal of the ENGAGE project is the deployment and use of an advanced service infrastructure, incorporating distributed and diverse public sector information resources as well as data curation, semantic annotation and visualisation tools, capable of supporting scientific collaboration and governance-related research from multidisciplinary scientific communities, while also empowering the deployment of open government data towards citizens. More information can be found at www.engage-project.eu and www.engagedata.eu. The authors would like to thank their colleagues of the ENGAGE project for their input for this paper, although the views expressed are the views of the authors and not necessarily of the project.

O**ver the last** years, a large number of studies has shown that opening data by public and private organisations have considerable potential to provide researchers, citizens, companies and other stakeholders with many advantages, such as a growing economy by stimulating innovation and obtaining new insights in the public and private sector by creating new ways of understanding problems and interpreting data (for instance, Blakemore & Craglia, 2006; Charalabidis, Ntanos, & Lampathaki, 2011; Dekkers, Polman, Velde, & Vries, 2006; European_Commission, 2003, 2011b; Zhang, Dawes, & Sarkis, 2005).

In practice, the advantages of open data are not realised on a large scale yet, as there are many impediments for the provision and use of open data. Considerable impediments are related to the provision of data about the data – so-called metadata (for instance, Dawes, 2010a; Whitmore, 2012). Among many examples, understanding and interpreting Linked Open Data (LOD) is difficult (Klischewski, 2012), as information about the context of the data, the time span, the validity, the quality, the accuracy and the comparability is missing. Other impediments concerning metadata are that it is difficult to search through or browse LOD, because the metadata are often not structured (Dawes, 2010b) and not machine readable. In addition, LOD are provided in a language that the user may not understand. Metadata could solve these problems, as metadata may yield considerable benefits (Dawes & Helbig, 2010). For instance, metadata may help to create order in datasets by describing, classifying and organizing information (Duval, Hodgins, Sutton, & Weibel, 2002; Zuiderwijk, Jeffery, & Janssen, 2012). Metadata are viewed as key enablers for the effective use of LOD.

To achieve the advantages of metadata for LOD, LOD should not just be seen as a product, but as an on-going process in which data are published, found, used, linked, reused and discussed, which is here referred to as the open data process (Janssen & Zuiderwijk, 2012). Generally, the open data process can be divided in five basic steps (Zuiderwijk, Janssen, Choenni, Meijer, & Sheikh_Alibaks, 2013). First, data are produced, collected and integrated by public and private organizations to fulfil their tasks. Second, these organizations decide whether the data that they own are published on the internet. Data that are published are referred to as open data. Third, open data may be found by potential open data users and, fourth, they can be reused and redistributed. One way of reusing data is by linking them to other data, so that relationships with other data on the Web can be revealed (Berners-Lee, 2009). Data that are both open and linked are referred to as Linked Open Data (LOD). Open data obtain more value when they are linked compared to open data that are not linked (Berners-Lee, 2009). In the fifth step in the open data process, feedback information on the use of the data is provided to the organizations that produced the data, so that these data may be used to improve work processes, such as policy making processes of public bodies. These five steps describe the high-level open data process.

However, in the open data process very little attention is paid to metadata. For instance, administrations express reluctance to publish metadata (ISA_Interoperability_Solutions_for_European_Public_Administrations, 2011; Schuurman, Deshpande, & Allen, 2008) and often publish insufficient metadata (Nichols, Twidale, & Cunningham, 2012; Tenopir et al., 2011; Xiong, Hu, Li, Tang, & Fan, 2011), there are no commonly agreed metadata (European_Commission, 2011a) and metadata are poorly documented (Conradie & Choenni, 2012). These impediments may result in ambiguous semantics of the data (Conradie & Choenni, 2012) and unawareness of the datasets by users (Schuurman, et al., 2008). Furthermore, metadata have received scant attention in research. One reason for this is that the benefits of metadata might not be clear at once and another reason is that it is often not clear how metadata can be used in LOD architectures.

This paper aims to determine the potential of metadata and its value for LOD users and publishers. Based on the outcomes, a meta-data model and an architecture is developed. Although we refer to this architecture with the term *metadata architecture* and this suggests that we will provide an architecture about metadata, we aim to provide an architecture for the use of LOD, which uses metadata for this purpose. In the following section, the research approach is presented. Further, information about metadata and an overview of advantages and disadvantages of metadata derived from literature is presented. Thereafter, metadata requirements for LOD architectures are derived from literature, workshops and a questionnaire. Subsequently, we describe an LOD architecture that meets the requirements and elements and we illustrate this architecture by showing illustrative examples of the application of this architecture in the ENGAGE LOD infrastructure.

Parts of this paper were published in the proceedings of the International Conference for E-Democracy and Open Government 2012 (Zuiderwijk, et al., 2012).

1. Research Approach

This paper aims to determine the potential of metadata and its value for LOD users and publishers and gives directions to release this potential by proposing an architecture. We start by creating a literature overview to determine what metadata are and which benefits they can provide. From this literature overview, we derive a definition of metadata and an initial list of advantages and disadvantages of metadata. In addition, requirements on metadata for use in the open data process are derived from literature, workshops and a questionnaire. On the basis of the requirements, a metadata architecture is developed. We opted for using various sources, as this is expected to provide a more comprehensive overview of metadata requirements than a single source. The approach of the literature overview, workshops and questionnaire is as follows.

1.1. Literature Overview

The literature overview is created by searching for journal papers, conference papers, books, governmental and non-governmental reports and other information in various databases, including Science Direct, Scopus and Google Scholar. Keywords that were used during this search were combinations of the terms metadata, meta data, advantages, benefits, disadvantages, problems and requirements. In total, approximately 5088 documents were found in the Science Direct database, 18.290 in Scopus and 1.960.000 in Google Scholar. There might be an overlap between those documents. The documents were filtered by searching for metadata advantages and disadvantages of and requirements for the use and provision of metadata for LOD. Most of the obtained documents appeared not to be useful, as they did not describe metadata advantages or requirements for the open data process. In total, 27 publications were selected that were relevant and from these an overview of the advantages and disadvantages of and requirements for metadata in the open data process was created. The relevance of the hits was determined by the search machines and by scanning the titles and abstracts of the documents.

1.2. Workshops

To broaden our knowledge about the potential use of metadata to support the open data process, we conducted four workshops at different international events. The workshops aimed at engaging a broad range of open data users, as different users are expected to mention different use possibilities and different advantages and disadvantages of the use of metadata. Multiple workshops were conducted in different countries so that a large number of open data users living in various countries would be reached, which decreases the risk on invalid or country specific conclusions. Table 1 shows the workshops that were organized.

Table 1: Organized workshops to obtain information about metadata benefits and requirements for the open data process.

Conference and workshop title	Location and date	Participants
1) International Conference for E-Democracy and Open Government (CeDEM12), “Open Linked governmental data for citizen engagement - A workshop about the benefits and restrictions of open linked governmental data and the role of metadata in citizen engagement” (90 minutes).	Danube University of Krems, Austria. May 4, 2012	17. Mainly civil servants, academic researchers, students.
2) Annual International Conference on Digital Government Research (DG.O2012), “Linking	Robert H. Smith School of Business, University of	26 (after the break 22). Mainly researchers (universities,

open data - Challenges and Solutions" (half day)	Maryland, USA. June 4, 2012	government, other).
3) Samos 2012 Summit on Open Data for Governance, Industry and Society (Samos Summit), "Open Data Requirements" (90 minutes)	University of the Aegean in Samos, Greece. July 3, 2012	16. Mainly students, academic researchers.
4) IFIP - Electronic Government Conference (IFIP EGOV 2012), "A workshop about using open public sector data: The ENGAGE project" (half day)	University of Agder, Norway. September 3, 2012	12 (after the break 10). Mainly researchers (universities), civil servants, companies.

Table 1 shows that a broad range of open data users coming from different countries was consulted during the workshops, including academic researchers, non-academic researchers, civil servants, companies and students. We expect that these users had different perspectives on metadata. However, as the number of participants might not be large enough to generalize the findings, we conducted a questionnaire among open data users to complement the results of the literature overview and workshops.

1.3. Questionnaire

Whereas the workshops had an explorative nature, a questionnaire about the use of open public sector data was developed to generalize the findings. The questionnaire aimed to gain information about the state of the art of using open public sector data in general and contained questions about the background of open data users, their use of open public sector data, metadata and statements about using open public sector data. In this paper, we concentrate on the part of the questionnaire that concerns metadata. The term metadata was defined as data about data and a number of examples was provided. The questions that were asked about metadata are as follows.

- To which extent are you currently able to 1) find data by the use of metadata, 2) download supplementary open public sector data (e.g. metadata) and 3) process data by linking metadata?
- To which extent do you find 1) finding data by the use of metadata, 2) downloading supplementary open public sector data (e.g. metadata) and 3) processing data by linking metadata useful for your use of open public sector data?
- Do you currently use metadata in the context of your work or for other activities?
- When you use metadata for open public sector data in your current practice, how often do you personally obtain the benefits 1) metadata can make reusing data easier, 2) metadata can make interpretation of data easier, 3) metadata can make searching and browsing data easier and 4) metadata can make linking data easier?
- When you use metadata for open public sector data in your current practice, how often do you personally notice the problems 1) insufficient metadata and therefore difficult to interpret the data, 2) insufficient data about the data quality, 3) insufficient metadata about data gathering and measuring and 4) metadata have no structure and are therefore difficult to search and browse?
- Which metadata would you like to use when you use (e.g. search, browse, retrieve and evaluate) open public sector data? (users could tick any type of metadata from a large list)

Before the questionnaire was published, it was evaluated by twelve persons. After processing the comments of the evaluators, a final version of the questionnaire was published on the internet in the beginning of April 2012. The questionnaire focused on users and potential users of open data (e.g. citizens, researchers, civil servants, developers, companies, journalists and archivists). In this paper we only use the results of the respondents who were actual users of open data. All fields of research were included, although the questionnaire mainly focused on researchers and citizens coming from social sciences and humanities. The questionnaire was conducted in several countries, including the Netherlands, Germany, Greece, Austria, Norway, Spain, the United Kingdom and the United States of America.

The respondents were informed about the confidential treatment of the information that was provided by their participation in the questionnaire. For actual users of open data, the questionnaire consisted of 19 to 23 questions, depending on whether the respondent used metadata and the field he mainly worked in. Completing the questionnaire took about 15 to 25 minutes.

To obtain a large user base, the questionnaire was disseminated as broad as possible. An online version as well as a paper version were disseminated. The URL to the online questionnaire was:

- Sent to e-mail lists of conferences and readers were asked to fill out the questionnaire;
- Included on the ENGAGE project website;
- Sent to researchers via LinkedIn and these researchers were asked to fill out the online questionnaire;
- Sent to people directly (for instance via the contact list of the ENGAGE project);
- Sent to organizations that employ researchers that probably work with open data;
- Sent to persons working for open data platforms, who were asked to put the link to the questionnaire on their website or in a newsletter (for instance the EPSI-platform, the website of the Dutch Data Archiving and Networked Services and Dutch governmental open data websites).

The paper version of the questionnaire was used for the following purposes:

- During workshops that were organized (see the section below). Workshop participants were asked to fill out the questionnaire and discuss it with each other and the workshop organizers. The discussion provided information about how reliable the answers to certain questions were.
- The questionnaire was printed and handed out to conference participants.

The questionnaire was disseminated between April 2012 and September 2012. During this period, 307 people filled out a part of the questionnaire and approximately 50 per cent of these persons filled out all questions.

1.4. Developing the ENGAGE Metadata Architecture

The literature overview, workshops and questionnaire resulted in a set of requirements and possible benefits of the use of metadata for LOD architectures. The development of the ENGAGE metadata architecture is based on these requirements. The development consisted of reviewing associated work in e-Infrastructure projects of relevance, such as those on the the European Strategy Forum on Research Infrastructures (ESFRI roadmap, see http://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri) and projects supporting it, such as the European Data Infrastructure (EUDAT, see http://www.eudat.eu/), Common Language Resources and Technology Infrastructures (CLARIN, see http://www.clarin.eu/external/), the Council of European Social Science Data Archives (CESSDA, see http://www.cessda.org/) and the Digital Research Infrastructure for the Arts and Humanities (DARIAH, see http://www.dariah.eu/). These projects are of immediate relevance, because they cover various aspects of Information and Communication Technologies (ICT) and social science.

In addition, we took account of the architecture in different areas of provision of research datasets to end-users, such as the European Plate Observing System (EPOS, see http://www.epos-eu.org/), and we designed the architecture around commonly accepted and widely-used standards for metadata and vocabularies. By using this approach, ENGAGE will have the widest possible take-up and acceptance, the maximum ability to interoperate with other portals to datasets and minimal maintenance effort.

2. The Potential of Metadata

In this section we define metadata and present an overview of advantages of metadata derived from literature.

2.1. Use of Meta-data for LOD

In the introduction we postulated that the open data process could be improved by providing considerable structured metadata. Various definitions of metadata are available. For instance, the National Information Standards Organization defines metadata as “structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource” (2004, p. 1). Gilliland (2008, http://www.getty.edu) states that metadata in the computing and information systems discipline refers to “a suite of industry or disciplinary standards as well as additional internal and external documentation and other data necessary for the identification, representation, interoperability, technical management, performance, and use of data contained in an information system”. A comparison of twenty-seven definitions of metadata referred to by Ma (2006, p. 4) resulted in the definition of metadata as “structured, encoded data that describe characteristics of information bearing entities to aid in the identification, discovery, assessment, and management of the described entities”. Often, metadata are simply defined as data about data (Jeffery, 2000; Schuurman, et al., 2008) or information about information (National_Information_Standards_Organization, 2004).

In the information system context, metadata may be classified (Jeffery, 2000) into schema metadata (which control the integrity of the described data), navigational metadata (which provide the access path to the data) and associative metadata divided into descriptive, restrictive and supportive categories. In the ideal situation, different types of metadata for LOD are provided, including the following types of metadata.

- Discovery (flat) metadata are descriptive and navigational and allow for the discovery of relevant open data by browsing or query. Examples of discovery metadata include data about the identifier, title, creator, publisher, country, source, type, format, language, sector, subjects, keywords, relative information system, validity date (from – to), audience, legal framework, status, relevant resources and linked data sets. Discovery metadata may be described by the metadata model Dublin Core (DC) (Dublin_Core_Metadata_Initiative, 2010), the e-Government Metadata Standard (e-GMS) (ESD_Standards, 2004), the platform Comprehensive Knowledge Archive Network (CKAN) (Open_Knowledge_Foundation, 2007) or similar ‘flat’ metadata.
- Contextual metadata are descriptive, restrictive and navigational. They allow for rich information on persons, organisations, projects, publications and many other aspects associated with the dataset. In addition, contextual metadata provide interoperation by ingesting and generating among common metadata formats used in open data. Contextual metadata are data about organizations, persons, projects, funding, facilities, equipment, services and pointers to detailed metadata. Contextual metadata may be described by the metadata model Common European Research Information Format (CERIF) (EuroCRIS, 2010). The European Union (EU) recommends CERIF to its member states. As CERIF is highly structured and flexible (Debruyne, Leenheer, Spyns, Grootel, & Christiaens, 2011), it allows for temporally defined role-based relationships between instances of entities.

- Detailed metadata cover schema metadata plus additional metadata to assure quality such as constraints on attribute values and values of one attribute conditional on another. Detailed metadata are usually specific to a domain (e.g. healthcare) or to a specific dataset. Detailed metadata can be data about the quality (accuracy, precision, calibration and other parameters (Charalabidis, et al., 2011) and domain or dataset-specific parameters that are used by software accessing and processing the dataset. CERIF points to the detailed metadata for each dataset instance that is contextually described in CERIF.

LOD is a relatively new field and the metadata models developed so far are ‘flat’ and therefore lack the required richness. LOD that are used by research communities are often context-specific with regard to time, application, provenance and nature. However, the original context of the dataset is often removed or not added (Charalabidis, et al., 2011). Although for some datasets with LOD detailed metadata are available, usually there are few and insufficient ways of managing metadata and interpretations of LOD (Tenopir, et al., 2011). For instance, there is no agreement about which metadata should be published for LOD (European_Commission, 2011a), which may result in numerous different ways of publishing LOD metadata. Another example is that metadata about the quality of the data, the way that the data were gathered and measured and domain-specific information are missing, which may result in different interpretations of one dataset (Zuiderwijk, et al., 2013). Moreover, metadata are provided in various forms, ranging from simple text files to easily reusable formats (Schuurman, et al., 2008).

Adding metadata is often viewed as an additional activity that only consumes resources. One reason for this is that those who might benefit from metadata are not the same persons or organisations as those who can provide the metadata. Furthermore, tools for preparing metadata are often unsatisfactory (Tenopir, et al., 2011). The existing meta-models provide limited value and these need to be extended to support the open data process.

In conclusion, our literature review shows that the current metadata provision is insufficient and especially contextual metadata are lacking. The current situation is far from the ideal situation, in which discovery (flat) metadata, contextual metadata and detailed metadata are all provided extensively to annotate and valorise datasets. Contextual metadata are critically important for the successful (re)use of LOD and not only encode the discovery metadata, but also provide pathways into more detailed metadata for particular datasets necessary for their (re)use.

2.2. Advantages and Disadvantages of Metadata

In this section an overview is provided of the advantages and disadvantages of metadata. Table 2 shows the advantages of different types of metadata and thereafter Table 3 shows the disadvantages of metadata. The overview of advantages is derived from literature and divided into five categories: accessibility of data, discovery of data, interpretation of data, linking data and other advantages.

Table 2: Overview of advantages of metadata, derived from literature.

	Benefits	Sources
Accessibility of data	1. Metadata improve storing and preservation of LOD, so that knowledge that can be derived from data become more explicit in a good state of preservation and are accessible in the future.	(King, Liakata, Lu, Oliver, & Soldatova, 2011; National_Information_Standards_Organization, 2004; Taylor, 2003; Tenopir, et al., 2011)
	2. Metadata improve the accessibility of LOD for others than the creator of the	(Bertot, Jaeger, Shuler, Simmons, & Grimes, 2009; Duval, et al.,

	data by describing, locating and retrieving the data efficiently.	2002; Joorabchi & Mahdi, 2011; Pallickara, Pallickara, & Zupanski, 2012; Tenopir, et al., 2011; United_Nations_Statistical_Commission_and_Economic_Commission_for_Europe, 2000)
Discovery of data	3. Metadata improve the ability to find LOD and the chances to be found by describing content and becoming searchable. Especially LOD that do not consist of text, such as images or sculptures, benefit from becoming searchable by adding metadata.	(Borgman, 2007; Jeffery, 2000; McGovern, 2001; National_Information_Standards_Organization, 2004; Schuurman, et al., 2008)
	4. Metadata make potential users aware of the existence of certain datasets.	(Schuurman, et al., 2008)
Interpretation of data	5. Metadata make sense of LOD by creating order within datasets by describing, classifying and organizing information.	(Berners-Lee, 2009; Duval, et al., 2002; National_Information_Standards_Organization, 2004)
	6. Metadata improve easily analysing, finding patterns, comparing, reproducing and finding inconsistencies in LOD. Making metadata available promotes the interchange of results and of the methods that are used to obtain these results.	(King, et al., 2011; Taylor, 2003; United_Nations_Statistical_Commission_and_Economic_Commission_for_Europe, 2000).
	7. Metadata provide a context for using data. Metadata improve chances of a correct interpretation of LOD and distilling knowledge from them. Metadata make potential reusers aware of the purpose or semantics of the data.	(Foulonneau & Cole, 2005; Jeffery, 2000; Schuurman, et al., 2008; United_Nations_Statistical_Commission_and_Economic_Commission_for_Europe, 2000; Vardaki, Papageorgiou, & Pentaris, 2009)
	8. Metadata may make it possible to assess and rank the quality and reliability of LOD. Therefore, the extensive use of metadata may lead to better decision making.	(Dawes, 2010b; Hönle, Käppeler, Nicklas, Schwarz, & Grossmann, 2005; Rahm & Hai_Do, 2000; Taylor, 2003)
	9. Metadata allow bringing similar resources together and distinguishing dissimilar resources.	(National_Information_Standards_Organization, 2004)
	10. Metadata provide a link between the creator of the data and the person who reuses these data.	(Taylor, 2003)
	11. Metadata may improve visualizing LOD. Metadata can, for instance, improve accuracy of mapping.	(Park, Kim, Seo, & Kim, 2011)

	12. Metadata enable detecting changes of LOD and therefore they can help in version management.	(Liu & Li, 2011; Sen, 2004)
	13. Metadata allow a dataset to be understood by both humans and machines in ways that promote interoperability, so that it enables multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality.	(National_Information_Standards _Organization, 2004)
	14. Metadata allow giving location information.	(National_Information_Standards _Organization, 2004)
	15. Metadata may stimulate collaboration by assisting in multilinguality and multimedia representations.	(Hüner, Otto, & Österle, 2011; Jeffery, 2000)
Linking data	16. Metadata make linking data easier. For instance, when data are provided about the accuracy, the precision and the calibration of a certain variable, these data can be used to decide whether or not to use this variable for the linking of data to other data.	(Rahm & Hai_Do, 2000; United_Nations_Statistical_Commission_and_Economic_Commission_for_Europe, 2000)
	17. Metadata facilitate legacy resource integration.	(National_Information_Standards _Organization, 2004)
	18. Metadata are essential for integrating and linking data from heterogeneous sources.	(Jeffery, 2000)
Other advantages	19. Metadata avoid unnecessary duplication of LOD, because similarities between data can be found more easily when extensive metadata are provided.	(King, et al., 2011)
	20. Metadata can increase the visibility of researchers and hereby stimulate collaboration among researchers from various organizations.	(Nonthakarn & Wuwongse, 2012)
	21. Metadata can help to create and maintain a common understanding of business objects and business processes, so that errors in automated activities and waiting times between activities can be minimized.	(Hüner, et al., 2011; Vardaki, et al., 2009)

Despite the many benefits there are a number of drawbacks of metadata. Table 3 shows the disadvantages of metadata, divided in the categories costs of metadata and interpretation of metadata. The overview is derived from literature.

Table 3: Overview of disadvantages of metadata, derived from literature.

	Disadvantages	Sources
Costs of metadata	Metadata may be sensitive and may be spread with the data unwillingly.	(Castiglione, De_Santis, & Soriente, 2007)
	Adding metadata is very time-consuming, as metadata operations consume over 60 per cent of the operations in typical workloads.	(Xiong, et al., 2011)
	Requires high investments and is costly.	(Duval, et al., 2002; Vardaki, et al., 2009)
Interpretation of data	The provision of considerable metadata makes it difficult to create consistency between metadata.	(Duval, et al., 2002)
	When metadata contain assumptions for the use of open data, they could point at certain choices and interpretations. This may unconsciously exclude certain ways of reusing data.	(Zuiderwijk, et al., 2013)

Table 2 shows that the use of metadata provides many benefits. On the other hand, Table 3 shows that some studies show disadvantages of the use of metadata for LOD. Adding metadata consumes resources and might require changes in the daily processes, whereas they can have a significant contribution for the adoption by users. Both the advantages and the disadvantages need to be considered in developing LOD architectures.

3. A Metadata Architecture for LOD

In this section, we describe which metadata requirements should be defined for LOD architectures to achieve the listed advantages and to avoid the disadvantages. Subsequently, a metadata model and its related architecture are presented. Finally, the architecture is illustrated by showing its application on the ENGAGE LOD platform.

3.1. Metadata Requirements for LOD Architectures

On the basis of Table 2 and 3, the workshops, the questionnaire and our background knowledge, we identified requirements for LOD architectures. We clustered the requirements into five categories based on the categorization of the advantages and disadvantages of metadata in section 3.2. The categories of accessibility and discovery are combined, as the requirements for those categories are overlapping. The category of sustainability is added, which refers to the ability of the metadata architecture to adapt to future changes and be usable on the long term by being flexible and changeable. Although benefits of metadata with regard to sustainability were not literally mentioned in the literature, several other benefits point at this category. For instance, metadata could enable long term sustainability by providing insight in the context and the quality of the data. Further, sustainability is of significant importance to create uptake of LOD, because potential users of LOD do not want to learn how to work with a metadata architecture that they can only use temporarily. All other advantage and disadvantage categories were adopted for the overview of metadata requirements.

Table 4: Overview of metadata requirements for LOD architectures.

	Requirement
Accessibility and discovery of data	1. It should be easy to add metadata and to use this metadata model in any other way for the provision of metadata.
	2. The metadata should be easily discovered by be easily accessible and highly visible.
	3. The metadata should provide a LOD representation of the metadata for searching, browsing and query.
	4. Downloading metadata should be enabled.
	5. The metadata should support multiple languages.
	6. Version management recorded in metadata should be clear to get insight in modifications made to the data, to correct errors that were made and to recover previous states of the data to stimulate storage and preservation.
Interpretation of data	7. Provide sufficient metadata fields to interpret the data (discovery, contextual and detailed metadata), including metadata about the (legal) context, gathering and measuring the data, the quality of the data (preferably standardized, including accuracy, the precision and the calibration of a certain variable) and persons that can be contacted about the data, such as the creator or provider of the dataset.
	8. Metadata should be consistent among different sources.
	9. Metadata should be neutral in the way that they do not point at wrongful choices and interpretations.
	10. Provide the possibility to visualize metadata.
Linking of data	11. Metadata should interconvert common metadata formats used for LOD.
	12. The metadata should maintain the capabilities of conventional information systems with structured query including convenient primitive operations.
	13. The metadata should be structured, standardized and machine readable.
	14. It should be possible to link metadata of different datasets. For this purpose, metadata about the accuracy, the precision and the calibration of a certain variable are needed.
Sustainability	15. Metadata models must allow for extensions and different types of metadata in different domains.
	16. The metadata model should be flexible.
Costs of metadata	17. The costs and investments of providing metadata should be kept as low as possible.
	18. The time-consumption for providing metadata should be kept as low as possible.

Table 4 shows a large number of requirements for LOD architectures. The requirements are described at a high-level of abstraction and are general. In our interviews it was found that the interviewees had a general idea about the requirements, but did not express them in detail. In the following section we describe how the requirements that were enumerated in Table 4 can be met by using a metadata model that uses standardization and harmonization techniques.

3.2. Metadata and Resource Description Framework (RDF)-Based Architecture

In this subsection we discuss the rationale for the design decisions taken. To meet the requirements, a metadata model was developed driven by our insights from literature and experiences. As shown in Figure 1, a three-layer structure for metadata is used:

1. discovery (flat) metadata;
2. contextual metadata;
3. detailed metadata.

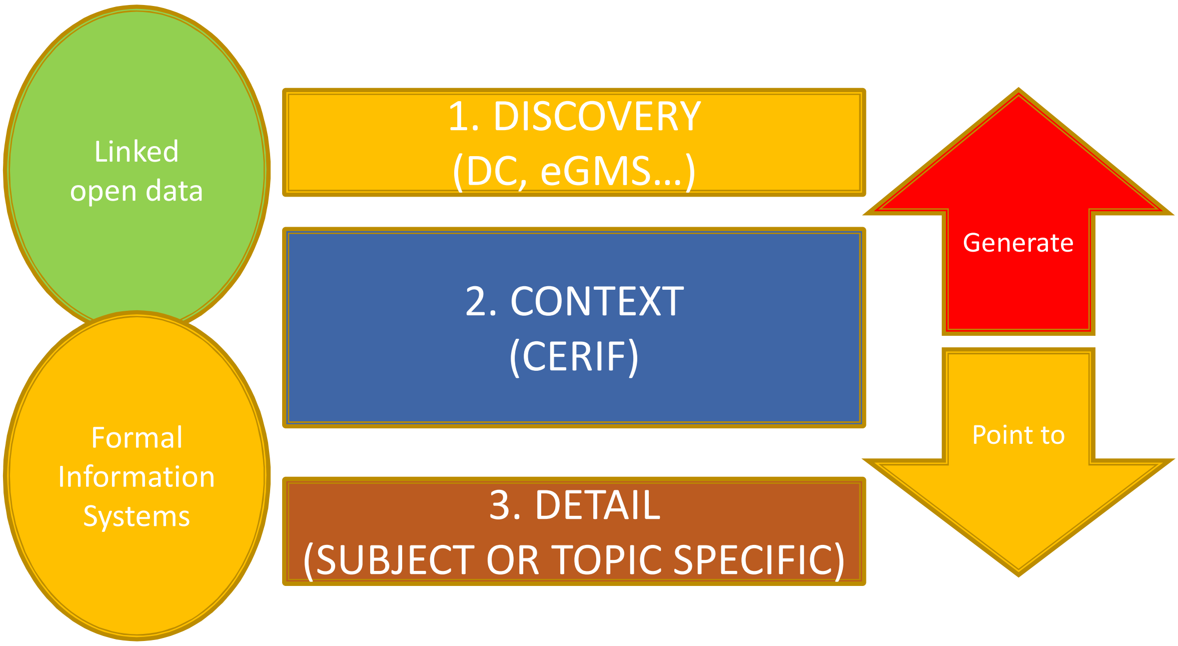

Figure 1: The three-layer metadata architecture

Of these layers, discovery (flat) metadata (1) allows discovery of relevant LOD by browsing or query, determining any high-level restrictions on usage. The layer of contextual metadata (2) allows rich information on persons, organisations, projects, publications and many other aspects associated with the dataset, including conditions of use. Furthermore, the layer of contextual metadata provides interoperation (by ingesting and generating) among common metadata formats used in PSI and from the contextual metadata we generate the discovery metadata to ensure congruence. Finally, the layer of detailed metadata (3) is usually specific to a domain, such as healthcare or crime, or even to a specific dataset. Examples include the European Commission INSPIRE Data portal (2012) or the Core Scientific Metadata Model (CSMD) (Matthews et al., 2010).

In this three layer environment, the layer of discovery metadata may be Dublin Core (DC) (Dublin_Core_Metadata_Initiative, 2010), the e-Government Metadata Standard (e-GMS) (ESD_Standards, 2004), the Comprehensive Knowledge Archive Network (CKAN) (Open_Knowledge_Foundation, 2007) or similar ‘flat’ metadata. These ‘flat’ metadata standards used in the discovery level have the advantage of simplicity, as it allows easy linkage of open datasets. The green oval in the top left corner of Figure 1 refers to this characteristic of discovery metadata.

However, the flat metadata standards used in the discovery layer have the disadvantage that they do not represent accurately the 'world of interest'. Although they do enable the easy linkage of large numbers of datasets, they insufficiently describe the relationships between those datasets. In particular, the syntax of flat metadata standards is often insufficiently formal, the semantics presented are rudimentary, they do not handle well multilinguality, they do not respect referential integrity and they do not handle well temporal relationships. All these disadvantages refer to very important aspects, as they could stimulate and support the reuse of open data. For instance, if the semantics are rudimentary, this might result in not finding the suitable dataset by a potential user, as these data might be described with terms that the potential user does not expect. Moreover, if datasets do not handle temporal relationships, the reuse of different versions of one dataset becomes very difficult, for instance, when a data provider regularly updates a dataset or when a data user uploads a cleaned or extended dataset.

Because of these disadvantages of flat metadata standards, we chose to add a second metadata layer to the metadata architecture. This second layer, the layer of contextual metadata, uses CERIF (EuroCRIS, 2010). CERIF is the only model that offers highly structured relationships, allowing temporally defined role-based relationships between instances of entities. Moreover, CERIF is an EU recommendation to member states. Furthermore, it is adopted by several governments (for example the United Kingdom, Norway, Denmark, Sweden, Slovakia, Slovenia, Ireland and the Netherlands) and by European institutions including European Research Council (ERC) and European Science Foundation (ESF). This ensures a large user-base and much knowledge is available. Moreover, CERIF is maintained by an independent organization called euroCRIS (www.eurocris.org), in this way ensuring continuity and adoption to changing needs.

Figure 1 shows that the layer of contextual metadata could generate discovery metadata. This is because the flat metadata models are proper subsets (i.e. are subsumed by) CERIF. In addition, it points to detailed metadata, which is the third layer of our metadata architecture. The layer of detailed metadata is added, because many datasets can be described by metadata that are specific for a certain domain or even for a certain dataset. The layer of detailed metadata allows the provision of such domain or dataset specific metadata in a formalized way.

Instead of following the architectural design that was described in the previous paragraphs, another design option considered was to only use the Semantic Web / LOD environment. This option is usually the one preferred for open public sector data, as this allows simple linkages between datasets, in this way being very flexible. In addition, it allows for the provision of simple metadata with a low effort threshold. Largely driven by enthusiasm from Web Science (WebScience_Trust, 2011) and World Wide Web Consortium (W3C) sources (World_Wide_Web_Consortium, 2011) – including Tim Berners-Lee and Nigel Shadbolt – governments have been persuaded that it is easy to make LOD available by this mechanism. Although this is useful for bottom-up linking of sources, the following aspects that are derived from a deeper analysis of the requirements described above challenge this view.

1. Firstly, the simplicity of the Semantic Web / LOD environment comes at the expense of not being able to add constrains and therefore lacks integrity. The flexibility allows adding similar metadata in different ways, which makes the resulting model complex and which makes it difficult to compare datasets.
2. Secondly, although access to metadata via a portal followed by 'point, click, download' to receive the dataset is easy, what is then done with the dataset? The end-user has to have adequate knowledge of the dataset and the metadata typically found at PSI websites (such as DC, CKAN or eGMS) is usually insufficient to provide this knowledge. Without the three-level metadata architecture there will be no standard format to add relevant information for discovery and about the context and details. Data providers will not be motivated to include these data which can have a profound negative effect on their use. Users will not know what the quality of the data is.

3. Thirdly, representation of the metadata using the Resource Description Framework (RDF) has intrinsic problems. RDF has the syntax of a simple triple <subject><link><object>, typically instantiated with Uniform Resource Identifiers (URIs). Although attributes may be added, this is already one level of indirection / complexity, and the simple <link> cannot adequately represent temporal or geospatial relationships (Perry, Sheth, & Jain, 2009) requiring multiple linked RDF statements. An example would be the representation of the fact 'between 1 January 2008 00:00 hours and 31 December 2013 24:00 hours the organisation Ministry of Health produced the product known as DatasetX'. Here the date/time start and end are obvious, the role is produced, the subject is the Ministry of health and the object DatasetX. A form of formal representation is {Ministry of Health}{<produced><200801010000><201312312400>}{DatasetX}, which is a triple of subject, relationship, object each enclosed thus: {..} but where the relationship has role and temporal duration with elements enclosed thus: <…>. Implementations of CERIF usually utilise relational database technology (although it is implementation environment agnostic) and so in this environment we provide interconversion between a relational representation and a RDF-representation (World Wide Web Consortium, 2010) of the contextual metadata. This allows the environment to be utilised in a semantic web / LOD context with guaranteed integrity as well as in a typical information system context.
4. Fourthly, the usual query language over RDF triple stores is the SPARQL Protocol and RDF Query Language (SPARQL) (World_Wide_Web_Consortium, 2008). Typically SPARQL endpoints are provided for the syntax of typical queries, the parameters of which can be instantiated by the user. Just as RDF is insufficient to carry the sophisticated semantics of relationships between metadata elements characterizing the LOD, so SPARQL becomes unacceptably complex (Perry, et al., 2009). The complexity might result in not using relevant information and ultimately to not using open data that might be relevant.

Because of these limitations, we do not use the conventional Semantic Web / LOD environment directly, but generate it from the environment described below. The three-level metadata architecture that was put forward in Figure 1, characterised typically by relational database technology, provides a number of advantages when it is compared with the Semantic Web / LOD environment. More specifically, the metadata environment provides the following improved facilities.

1. Firstly, CERIF provides a much richer metadata than the standards used commonly with LOD and so improves greatly the experience of the end user (or the software) in processing the PSI datasets described by the enhanced metadata. Therefore, CERIF may be important in realizing the advantages of providing metadata that were mentioned in the literature review section.
2. Secondly, the representation of contextual metadata (CERIF) allows rich semantics to be represented simply over a formal syntax, thus making the PSI datasets understandable to the end user (or software) through the enhanced metadata and to making adding metadata clearer and simpler to the metadata provider. Because CERIF has a structure of <entity><relationship><entity> it has outwardly a similarity to RDF. However, CERIF provides this structure at the entity (object) level as well as the instance level – a characteristic of data models based on the Entity-Relationship model (ER) (Wikipedia, 2011b) or more accurately the Enhanced Entity-Relationship model (EER) (Wikipedia, 2011a) paradigm. Nonetheless, this similarity makes it easy to represent in RDF (subject to the problems discussed above under the heading metadata) metadata represented within CERIF. Similarly, metadata represented in RDF can, with some manipulation, be represented in CERIF. Thus, CERIF provides the 'exchange' format for metadata allowing not only interconversion of common metadata formats as mentioned before but also between information system (typically relational) and semantic web / LOD (typically RDF) representations. In this way, metadata can help in providing an understandable common interface to the user of the data.

3. Thirdly, the Structured Query Language (SQL) (Wikipedia, 2011d) usually presented to the end-user through an easy-to-use Query By Example (QBE) interface (Wikipedia, 2011c) has a simpler structure than SPARQL and includes convenient primitive operations for simple statistical calculations such as sum, count, average.

Whether these facilities indeed result in the expected advantages for the use of LOD, will be evaluated in section 4.4.

To summarize, although metadata require additional resources to add the information and they might be viewed as a top-down format which might constrain the flexibility, we opted for using advanced metadata based on CERIF in our architecture. Our focus is on the user and lowering the threshold for making use of open data. Although for data providers bottom-up linking might be a quick approach, it is an approach lacking integrity for the users. There will be a void in data necessary for the easy discovery, the understanding of the quality, the content and other detailed information, which are extremely relevant for the users. Without metadata, discovery might not even be possible and wrong use, interpretation and conclusions might be drawn. In addition, by representing metadata using RDF, generated from the underlying formal database or information system environment utilising CERIF, the necessary flexibility is created and the best of the bottom-up world of linking data within a top-down frame guided by metadata is created. In other words, whereas unconstrained bottom-up linking at a first glance seems to improve the flexibility to use LOD, it constrains the possibilities to make use of the data on the long run, resulting in less flexibility for the users. By having clear metadata available, users are empowered and are not constraint by difficult structures which might be hard to grasp and constrain them in finding and understanding datasets.

In summary, we choose an architecture combining the ‘best of both worlds’; the world of formalised metadata and the world of easy to use RDF. Because of the powerful expressive semantics over formal syntax of CERIF we can:

- Generate discovery metadata from CERIF;
- Interconvert common metadata formats used in open data using CERIF as the superset exchange mechanism;
- Provide a semantic web / LOD representation of the formalized metadata for browsing or query using SPARQL;
- While maintaining a conventional information systems capability with structured query including convenient primitive operations.

This combined architecture is ideal for both the end-user via a portal access and running software via a service Application Programming Interface (API) access. It provides convenient and easy LOD/semantic web browsing but based on formalised metadata with data integrity because it is generated from the underlying formal database or information system environment utilising CERIF. It is easy to ‘pass through’ the semantic web/LOD view and utilise full-scale data processing operations at the underlying ICT environment level. Furthermore, it helps in realising the benefits of providing metadata that were enumerated in the literature review section.

3.3. Implementing the Metadata Architecture

In the previous section, we described a metadata architecture that could provide many benefits. In this section we describe how the metadata architecture can be realized by designing an application architecture. The application architecture is aimed at combining the best of both worlds by combining metadata and RDF. The overview of the ENGAGE application architecture is presented in Figure 2 and shows how CERIF and RDF are linked to each other. In this way, the application architecture of the underlying formal database or information system environment utilising CERIF and using conventional relational database technology with integrity, and the outer environment of LOD/semantic web based on RDF, are coupled.

The three layer structure of the presented metadata model meets the requirements that were described in section 4.1. First, it provides metadata fields to interpret the data, including discovery, contextual and detailed metadata. The first layer of the metadata model, the discovery (flat) metadata, allows the discovery of relevant open data by browsing, searching or query. The second layer of contextual metadata allows rich information on many aspects associated with the dataset, including the required metadata fields about the context, quality and contact persons. The provision of certain contextual and detailed metadata, such as the geographical and temporal coverage of the dataset, allows for the visualization of metadata. Moreover, metadata provision in different languages is supported by the metadata model. In addition, the contextual metadata provides inoperability among common metadata formats used for open data. In this way, it also meets the requirements that the metadata should interconvert common metadata formats used for LOD, that it should maintain the capabilities of conventional information systems with structured query including convenient primitive operations and that it should be structured, standardized, machine-readable, a LOD representation, linkable and flexible. The third metadata layer, the detailed metadata, allows the provision of detailed and domain specific metadata.

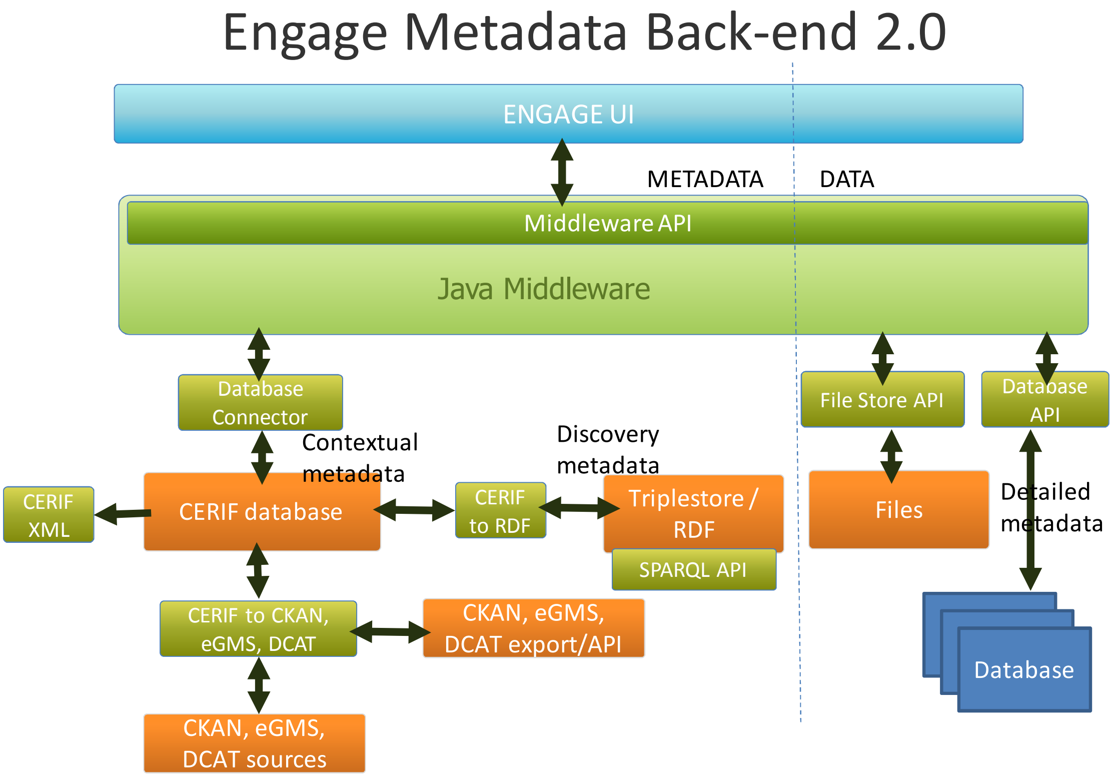

Figure 2. Concept of metadata architecture implementation in the ENGAGE infrastructure (EuroCRIS, 2011).

The foregoing shows that all requirements are met by the architecture. To summarize, the architecture makes it easy to add and use metadata. The metadata can easily be discovered and downloaded. In addition, the metadata architecture provides a flexible LOD representation of the metadata for searching, browsing and query and sufficient metadata fields to interpret and link the data and metadata (discovery, contextual and detailed metadata). Furthermore, the architecture allows for extensions and different types of metadata in different domains. It can support multiple languages. The metadata are structured, standardized and machine readable, they interconvert common metadata formats used for LOD, while they maintain the capabilities of conventional information systems with structured query including convenient primitive operations. Moreover, the formalized metadata stimulate the consistency and neutrality of metadata among different datasets and enable temporal version management and visualizations. Finally, the architecture keeps the costs, investments and time-consumption for adding metadata as low as possible.

These requirements stimulate the realisation of a large number of benefits, which will be illustrated in the following section.

3.4. Illustrating and Evaluating the Metadata Architecture Using a Scenario

To illustrate and evaluate how the requirements are met by the metadata architecture and how the benefits are achieved, a scenario for the use of an open data platform is given in this subsection (based on Charalabidis, 2012). This scenario is not derived from a real use case, but it aims to illustrate the most important functionalities that the ENGAGE platform provides and to evaluate whether the expected advantages originating from the metadata model are achieved and how they are realized. In the scenario description we show between brackets the realised benefits as they were presented in section 3.2.

Let us imagine that a researcher working for a university wants to perform research on European crime data and write a scientific paper about the way that crime has developed in the last decade. The researcher finds a URI for a dataset somewhere in the cloud. To analyse the dataset, he enters the ENGAGE platform (www.engagedata.eu) and pastes the URI of the dataset that he would like to investigate. The ENGAGE platform starts retrieving information from this URI. Additionally, it automatically fills all metadata fields, including the title, publisher, creator, classification code, person who uploads, subject, type, format, language, country, public sector domain, relative public service, relative information system, legal framework, scientific sector, scientific usage of resources, intended audience, keywords, type of source link, source link, type of resource link, resource link, relevant resources, linking status, linked datasets and visualisations. This process shows that metadata allow a dataset to be understood by both humans and machines in ways that promote interoperability (benefit 13).

Subsequently, the ENGAGE platform shows a report of the import. The report shows how many metadata attributes belonging to the dataset that the researcher wants to investigate the ENGAGE platform managed to import. Furthermore, it shows how many and which metadata attributes should mandatorily and non-mandatorily be filled to increase the validity of the dataset and to add it to the ENGAGE platform. The researcher adds two mandatory metadata fields that were not filled by the automatic metadata insertion and uploads the dataset that he wants to examine to www.engagedata.eu.

Now the researcher can start using the dataset. The automatically filled metadata have created order in the dataset by describing, classifying and organizing its information (benefit 5). Thus, the data can easily be reused and interpreted because sufficient metadata about the way that the data can be reused and interpreted are provided (benefit 6 and 7). The researcher cleans the dataset (i.e. he removes redundant data) and performs a statistical analysis on the ENGAGE platform. As the dataset contains geographical information, he is able to visualise a part of the dataset by showing it on a map (benefit 11 and 14). He uploads this derived dataset to the ENGAGE platform, just like he did with the previous dataset, so that the dataset is in a good state of preservation and is accessible in the future (benefit 1). The original and the derived dataset are now related to each other, so that other users of the ENGAGE platform can see which different versions of the dataset exist (benefit 12).

Thereafter, the researcher would like to relate this dataset to other datasets about crime developments in the last decade. He uses the advanced dataset search on the ENGAGE platform to find more European crime data. One of the advantages of the ENGAGE platform will be that it provides metadata about different aspects of the quality of the dataset (benefit 8). The ENGAGE platform enables searching through datasets of vast amounts of public sector datasets that are currently publicly available in flat metadata repositories, isolated data silos or linked data clouds and on the other with rich, highly structured metadata repositories from e-infrastructures, research data centres, statistical offices and data archives by providing community-driven services that allow the extraction of high re-use value out of open public data in a crowd-sourcing manner (benefit 2). Therefore, this metadata model is compatible and linkable with many PSI websites and national open data websites of many countries and has the capability of containing rich, contextual information with regard to the datasets (benefit 18). The metadata model is expected to provide European crime data that could satisfy the researcher and help him with linking the datasets (benefit 16) and writing his scientific paper (benefit 3 and 4). Moreover, the metadata search makes it possible to identify and distinguish similar sources and avoid the unnecessary duplication of the data (benefit 9 and 19).

Via the advanced search on the ENGAGE platform, the researcher may find some datasets that satisfy him. For instance, he finds crime data from the United Kingdom, the Netherlands, Germany, France and Greece. The metadata of these datasets are provided in various languages, so that he can easily understand them (benefit 16). The researcher downloads the datasets and analyses them. The datasets include license information, so that the researcher knows under which conditions he is allowed to reuse the dataset. However, he wants to analyse crime data from other countries as well. For this reason, he searches through the metadata on the ENGAGE platform and finds out which public sector organisations provide crime data in European countries. He contacts data publishers working for these public sector organisations via the ENGAGE platform and requests the data that he is looking for (benefit 10 and 20).

In addition, the researcher finds several user groups that discuss datasets on the ENGAGE platform. He sees that there is one group that consists of researchers, citizens, developers, journalists, public servants, businesses, and archivists who discuss crime data. Becoming a member of this group provides the researcher with the possibility to discuss his derived datasets and to ask questions about the way that crime data can be interpreted and compared across different European countries. Metadata in this way stimulate collaboration (benefit 15).

This scenario for the use of the ENGAGE platform showed the realisation of many benefits of metadata. It contains almost all the benefits mentioned in section 3.2. The scenario has shown that the metadata in our metadata model provide the benefits as follows. They:

- Improve the accessibility of data, as they:
    - improve storing and preservation of LOD
    - improve the accessibility of LOD for others than the creator of the data
- Improve the discovery of data, as they:
    - improve the ability to find LOD;
    - make potential users aware of the existence of certain datasets.
- Improve the interpretation of data, as they:
    - make sense of LOD by creating order within datasets by describing, classifying and organizing information;
    - improve easily analysing, finding patterns, comparing, reproducing and finding inconsistencies in LOD;
    - provide a context for using data;
    - make it possible to assess and rank the quality and reliability of LOD;
    - allow bringing similar resources together and distinguishing dissimilar resources;
    - provide a link between the creator and user of the data;
    - improve visualizing LOD;

  - improve version management;
  - allow a dataset to be understood by both humans and machines in ways that promote interoperability;
  - allow giving location information;
  - stimulate collaboration.
- Improve linking data, as they:
  - make linking data easier;
  - facilitate legacy resource integration;
  - support integrating and linking data from heterogeneous sources.
- Improve other advantages, as they:
  - avoid unnecessary duplication of LOD;
  - improve researcher visibility and collaboration.

The only benefit that was not immediately realised in this scenario is the benefit that metadata can help to create and maintain a common understanding of business objects and business processes (benefit 21). This is difficult to evaluate as it might vary per person based on his or her experiences, knowledge and other capabilities. However, the use of rich semantics in the contextual metadata layer, and the ability to crosswalk from one semantic term to another, certainly assists in this understanding. A benefit that we did not find in literature, but that was realised in the scenario is that metadata can stimulate international comparisons of data, as they support comparability and compatibility.

With this metadata model, all other benefits are realised, whereas the use of other metadata models is expected to have resulted in achieving less benefits.

4. Conclusions

LOD is a relatively new field and metadata models have not been developed yet. Adding metadata is often viewed as an additional activity that only consumes resources, whereas it can provide numerous benefits. Our literature review showed that metadata may improve storing, preservation, accessibility, visualisation and multilinguality of LOD, improve the ability to find and interpret LOD, increase awareness of the existence of certain LOD, create order within datasets, stimulate analysing, finding patterns, comparing, reproducing and finding inconsistencies in LOD, enable assessing and ranking the quality and reliability of LOD, make linking data easier, avoid the unnecessary duplication of LOD and improve visibility of and collaboration among researchers. Metadata may also have some disadvantages, such as the time-consumption, high investments and costs of adding and maintaining metadata and the inconsistency between metadata and assumptions and interpretations of metadata. Using metadata requires a trade-off between its advantages and disadvantages.

On the basis of the overview of advantages and disadvantages of metadata, an overview of eighteen requirements for metadata architectures was created, including making it easy to add, use, discover and download metadata, providing a flexible LOD representation of the metadata for searching, browsing and query, providing sufficient metadata fields to interpret and link the data and metadata (discovery, contextual and detailed metadata), allowing for extensions and different types of metadata in different domains, supporting multiple languages, providing structured, standardized and machine readable metadata, interconverting common metadata formats used for LOD, while maintaining the capabilities of conventional information systems with structured query including convenient primitive operations, stimulating consistency and neutrality of metadata among different datasets, enabling temporal version management and visualizations and keeping the costs, investments and time-consumption for adding metadata as low as possible.

To meet the requirements, a three-layer structure for metadata was used, including 1) discovery (flat) metadata, which allows discovery of relevant open data by browsing or query, 2) contextual metadata, which allows a) rich information on persons, organisations, projects, publications and many other aspects associated with the dataset, b) interoperation among common metadata formats used in PSI and from the contextual metadata we generate the discovery metadata to ensure congruence; 3) detailed metadata, which is usually specific to a domain or even to a dataset.

Given the potential of metadata, an infrastructure for LOD was developed to realize these advantages. We chose to use CERIF to implement the three-layer metadata architecture, because 1) it is the only model that offers highly structured relationships, allowing temporally defined role-based relationships between instances of entities, 2) it is an EU recommendation to member states, 3) it is adopted by several governments, and 4) it is maintained by an independent organization, in this way ensuring continuity and adoption to changing needs.

These formalised metadata were combined with the use of RDF. This combined architecture provides most benefits for both the end-user via a portal access and running software via a service API access. It provides convenient and easy LOD/semantic web browsing, but based on formalised metadata with data. It is easy to ‘pass through’ the semantic web/LOD view and utilise full-scale data processing operations at the inner environment level. Furthermore, it helps in realising the benefits of providing metadata that were enumerated in the literature review section. The latter is confirmed by the ENGAGE project, as the application of the proposed metadata architecture in the ENGAGE project showed that the metadata architecture could result in the realisation of all benefits that were found in literature.

References

Berners-Lee, T. (2009). Linked data. Retrieved October 11, 2012, from http://www.w3.org/DesignIssues/LinkedData.html

Bertot, J. C., Jaeger, P. T., Shuler, J. A., Simmons, S. N., & Grimes, J. M. (2009). Reconciling government documents and e-government: Government information in policy, librarianship, and education. *Government Information Quarterly, 26*(3), 433–436.

Blakemore, M., & Craglia, M. (2006). Access to Public-Sector Information in Europe: Policy, rights and obligations. *The Information Society, 22*(1), 13-24.

Borgman, C. L. (2007). *Scholarship in the Digital Age: information, infrastructure, and the internet*. Cambridge: MIT Press.

Castiglione, A., De_Santis, A., & Soriente, C. (2007). Taking advantages of a disadvantage: Digital forensics and steganography using document metadata. *The Journal of Systems and Software, 80*, 750-764.

Charalabidis, Y. (2012). On metadata for Open Data. Retrieved November 18, 2012, from http://www.slideshare.net/charalabidis/on-metadata-for-open-data?from=new_upload_email

Charalabidis, Y., Ntanos, E., & Lampathaki, F. (2011). An architectural framework for open governmental data for researchers and citizens. In M. Janssen, A. Macintosh, J. Scholl, E. Tambouris, M. Wimmer, H. d. Bruijn & Y. H. Tan (Eds.), *Electronic government and electronic participation joint proceedings of ongoing research and projects of IFIP EGOV and ePart 2011* (pp. 77-85). Delft

Conradie, P., & Choenni, S. (2012, October 22-25). *Exploring process barriers to release public sector information in local government.* Paper presented at the 6th international conference on theory and practice of electronic governance (ICEGOV), Albany, New York, United States of America.

Dawes, S. (2010a). *Information policy meta-principles: stewardship and usefulness.* Paper presented at the 43rd Hawaii International Conference on System Sciences, Grand Wailea, Maui, Hawaii.

Dawes, S. (2010b). Stewardship and usefulness: Policy principles for information-based transparency *Government Information Quarterly, 27*(4), 377–383.

Dawes, S., & Helbig, N. (2010). *Information strategies for open government: Challenges and prospects for deriving public value from government transparency.* Paper presented at the 9th International Conference on e-government (EGOV), Lausanne, Switzerland.

Debruyne, C., Leenheer, P., Spyns, P., Grootel, G., & Christiaens, S. (2011). Publishing Open Data and Services for the Flemish Research Information Space. In O. Troyer, C. Bauzer Medeiros, R. Billen, P. Hallot, A. Simitsis & H. Mingroot (Eds.), *Advances in Conceptual Modeling. Recent Developments and New Directions* (Vol. 6999, pp. 389-394): Springer Berlin Heidelberg.

Dekkers, M., Polman, F., Velde, R. t., & Vries, M. d. (2006). *MEPSIR. Measuring European Public Sector Information Resources*. Retrieved October 9, 2011, from http://ec.europa.eu/information_society/policy/psi/actions_eu/policy_actions/mepsir/index_en.htm.

Dublin_Core_Metadata_Initiative. (2010). Dublin Core Metadata Element Set, Version 1.1. Retrieved December 2, 2011, from http://dublincore.org/documents/dces/

Duval, E., Hodgins, W., Sutton, S., & Weibel, S. L. (2002). Metadata principles and practicalities. *D-lib magazine, 8*(4), unknown.

ESD_Standards. (2004). e-GMS-e-Government Metadata Standard version 3.0. Retrieved December 2, 2011, from http://www.esd.org.uk/standards/egms/

EuroCRIS. (2010). CERIF Releases Retrieved December 2, 2011, from http://www.eurocris.org/Index.php?page=CERIFreleases&t=1

EuroCRIS. (2011). ENGAGE metadata back-end 2.0.

European_Commission. (2003). Directive 2003/98/EC of the European Parliament and of the council of 17 November 2003 on the re-use of public sector information. Retrieved December 12, 2012, from http://ec.europa.eu/information_society/policy/psi/rules/eu/index_en.htm.

European_Commission. (2011a). *Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions. Open data. An engine for innovation, growth and transparent governance*. Brussels: European_Commission.

European_Commission. (2011b). *Digital agenda: Turning government data into gold*. Brussels: European_Commission.

European_Commission_INSPIRE_Dataportal. (2012). Welcome to the INSPIRE geoportal. Retrieved December 12, 2012, from http://inspire-geoportal.ec.europa.eu/

Foulonneau, M., & Cole, T. W. (2005). Strategies for reprocessing aggregated metadata. Retrieved December 12, 2012, from http://imlsdcc.grainger.uiuc.edu/docs/metadatareprocessing.pdf

Gilliland, A. J. (2008). Setting the stage. In M. Baca (Ed.), *Introduction to metadata. Version 3.0*. Los Angeles: Getty Research Institute.

Hönle, N., Käppeler, U.-P., Nicklas, D., Schwarz, T., & Grossmann, M. (2005). *Benefits of Integrating Meta Data into a Context Model.* Paper presented at the The third International Conference on Pervasive Computing and Communications Workshops.

Hüner, K., Otto, B., & Österle, H. (2011). Collaborative management of business metadata. *International Journal of Information Management, 31*, 366–373.

ISA_Interoperability_Solutions_for_European_Public_Administrations. (2011). *Towards open government metadata*. Brussels: ISA Interoperability Solutions for European Public Administrations.

Janssen, M., & Zuiderwijk, A. (2012). *Open data and transformational government.* Paper presented at the Transforming Government Workshop.

Jeffery, K. G. (2000). Metadata: The future of information systems. In J. Brinkkemper, E. Lindencrona & A. Sølvberg (Eds.), *Information Systems Engineering: State of the art and research themes*. London: Springer Verlag.

Joorabchi, A., & Mahdi, A. E. (2011). An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata. *Journal of Information Science, 37*(5), 499-514.

King, R. D., Liakata, M., Lu, C., Oliver, S. G., & Soldatova, L. N. (2011). On the formalization and reuse of scientific research. *Journal of the Royal Society Interface, 8*, 1440–1448.

Klischewski, R. (2012). *Identifying Informational Needs for Open Government: The case of Egypt*. Paper presented at the 45th Hawaii International Conference on System Sciences.

Liu, F., & Li, X. (2011). *Using Metadata to Maintain Link Integrity for Linked Data*. Paper presented at the Proceedings of the 2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing.

Ma, J. (2006). Managing metadata for digital projects. *Library Collections, Acquisitions, and Technical Services, 30*(1–2), 3-17.

Matthews, B., Sufi, S., Flannery, D., Lerusse, L., Griffin, T., Gleaves, M., et al. (2010). Using a core scientific metadata model in large-scale facilities. *The International Journal of Digital Curation, 1*(5), 106-118.

McGovern, G. (2001). Why metadata is important. Retrieved December 7, 2011, from http://www.gerrymcgovern.com/nt/2001/nt_2001_10_01_metadata.htm

National_Information_Standards_Organization. (2004). *Understanding metadata*. Bethesda: National Information Standards Organization Press.

Nichols, D. M., Twidale, M. B., & Cunningham, S. J. (2012). *Metadatapedia: a proposal for aggregating metadata on data archiving*. Paper presented at the Proceedings of the 2012 iConference.

Nonthakarn, C., & Wuwongse, V. (2012). Linked OpenScholar: A Researcher Network Using Linked Open Data. In H.-H. Chen & G. Chowdhury (Eds.), *The Outreach of Digital Libraries: A Globalized Resource Network* (Vol. 7634, pp. 325-328): Springer Berlin Heidelberg.

Open_Knowledge_Foundation. (2007). CKAN. Retrieved December 2, 2011, from http://ckan.org/

Pallickara, S. L., Pallickara, S., & Zupanski, M. (2012). Towards efficient data search and subsetting of large-scale atmospheric datasets. *Future Generation Computer Systems, 28*, 112–118.

Park, Y. R., Kim, H. H., Seo, H. J., & Kim, J. H. (2011). CDISC Transformer: a metadata-based transformation tool for clinical trial and research data into CDISC standards. *KSII Transactions on Internet and Information Systems, 5*(10), 1830-1840.

Perry, M., Sheth, A. P., & Jain, P. (2009). SPARQLST: Extending SPARQL to support spatiotemporal queries. Retrieved December 12, 2012, from http://knoesis.wright.edu/students/prateek/sparql-st-www09-tr.pdf

Rahm, E., & Hai_Do, H. (2000). Data Cleaning: Problems and Current Approaches. *IEEE Data Engineering Bulletin* Retrieved November 6, 2012, from http://dc-pubs.dbs.uni-leipzig.de/files/Rahm2000DataCleaningProblemsand.pdf

Schuurman, N., Deshpande, A., & Allen, D. (2008). Data integration across borders: a case study of the Abbotsford-Sumas aquifer (British Columbia/Washington State). *JAWRA Journal of the American Water Resources Association, 44*(4), 921-934.

Sen, A. (2004). Metadata management: past, present and future. *Decision Support Systems, 152*(37), 151-173.

Taylor, C. (2003). An introduction to metadata. Retrieved December 8, 2011, from http://www.library.uq.edu.au/papers/ctmeta4.html

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., et al. (2011). Data Sharing by Scientists: Practices and Perceptions. *PLoS ONE, 6*(6), e21101.

United_Nations_Statistical_Commission_and_Economic_Commission_for_Europe. (2000). Guidelines for statistical metadata on the internet. Retrieved December 8, 2011, from http://www.unece.org/fileadmin/DAM/stats/publications/metadata.pdf

Vardaki, M., Papageorgiou, H., & Pentaris, F. (2009). A statistical metadata model for clinical trials' data management. *Computer methods and programs in biomedicine, 9*(5), 129-145.

WebScience_Trust. (2011). Web Science. Retrieved December 2, 2011, from http://webscience.org/home.html

Whitmore, A. (2012). Extracting knowledge from U.S. department of defense freedom of information act requests with social media. *Government Information Quarterly, 29* (2), 151–157.

Wikipedia. (2011a). Enhanced Entity-Relationship Model. Retrieved December 2, 2011, from http://en.wikipedia.org/wiki/Enhanced_Entity-Relationship_Model

Wikipedia. (2011b). Entity-relationship model. Retrieved December 2, 2011, from http://en.wikipedia.org/wiki/Entity-relationship_model

Wikipedia. (2011c). Query by Example. Retrieved December 2, 2011, from http://en.wikipedia.org/wiki/Query_by_Example

Wikipedia. (2011d). SQL. Retrieved December 8, 2011, from http://en.wikipedia.org/wiki/SQL

World_Wide_Web_Consortium. (2008). SPARQL Query Language for RDF. Retrieved December 2, 2011, from http://www.w3.org/TR/rdf-sparql-query/

World_Wide_Web_Consortium. (2011). W3C. Retrieved December 2, 2011, from www.w3.org

Xiong, J., Hu, Y., Li, G., Tang, R., & Fan, Z. (2011). Metadata Distribution and Consistency Techniques for Large-Scale Cluster File Systems. *IEEE Transaction on parallel and distributed systems, 22*(5), 803-816.

Zhang, J., Dawes, S. S., & Sarkis, J. (2005). Exploring stakeholders' expectations of the benefits and barriers of e-government knowledge sharing. *Journal of Enterprise Information Management, 18* (5), 548-567.

Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., & Sheikh_Alibaks, R. (2013). Socio-technical impediments of open data. *Electronic Journal of eGovernment.*

Zuiderwijk, A., Jeffery, K., & Janssen, M. (2012, May 3-4). *The necessity of metadata for open linked data and its contribution to policy analyses.* Paper presented at the Conference on E-Democracy and Open Government (CeDEM12), Krems, Austria.

About the Authors

Anneke Zuiderwijk MSc

Anneke Zuiderwijk is a researcher in the Information and Communication Technology section of the Faculty of Technology, Policy, and Management at Delft University of Technology, the Netherlands. Her research concentrates on open linked data. More specifically, her research is focused on the development of a socio-technical infrastructure that coordinates opening data by organizations and using these data by different types of users. Furthermore, Anneke performs research on open data at the Research and Documentation Centre (WODC) of the Dutch Ministry of Security and Justice. She is involved in the FP7 ENGAGE project (An Infrastructure for Open, Linked Governmental Data Provision Towards Research Communities and Citizens). Anneke can be contacted via a.m.g.zuiderwijk-vaneijk@tudelft.nl.

Prof. Dr. Keith Jeffery

Keith Jeffery is currently Director IT and International Strategy at STFC (Science and Technology Facilities Council). Keith previously had operational responsibility for IT with 360,000 users, 1100 servers and 140 staff. Keith holds 3 honorary visiting professorships, is a Fellow of the Geological Society of London and the British Computer Society, is a Chartered Engineer and Chartered IT Professional and an Honorary Fellow of the Irish Computer Society. Keith is currently president of ERCIM and president of euroCRIS, and serves on international expert groups, conference boards and assessment panels.

Dr. Marijn Janssen

Marijn Janssen is an associate professor in the Information and Communication Technology section of the Faculty of Technology, Policy, and Management at Delft University of Technology, the Netherlands. His research is focused on the governance, design and orchestration of public-private service networks. Specifically, his research concerns the impact of and policy developments which fundamentally change such networks. He is also the director of both the interdisciplinary SEPAM and the Compliance Design & Management Master programmes. He serves on several editorial boards and published over 220 refereed publications. More information: www.tbm.tudelft.nl/marijnj, or contact him via m.f.w.h.a.janssen@tudelft.nl.