informationresearch

vol. 18 no. 3, September, 2013

Proceedings of the Eighth International Conference on Conceptions of Library and Information Science, Copenhagen, Denmark, 19-22 August, 2013

From many records to one graph: heterogeneity conflicts in the linked data restructuring cycle

Kim Tallerås,
Oslo and Akershus University College, Department of Archive Studies, Library and Information Science, Lilleborggatan. 3, 04480 OSLO, Norway

Abstract

Introduction. *During the last couple of years the library community has developed a number of comprehensive metadata standardization projects inspired by the idea of linked data, such as the BIBFRAME model. Linked data is a set of best practice principles of publishing and exposing data on the Web utilizing a graph based data model powered with semantics and cross-domain relationships. In the light of traditional metadata practices of libraries the best practices of linked data imply a restructuring process from a collection of semi-structured bibliographic records to a semantic graph of unambiguously defined entities. A successful interlinking of entities in this graph to entities in external data sets requires a minimum level of semantic interoperability.*
Method *The examination is carried out through a review of the relevant research within the field and of the essential documents that describe the key concepts.*
Analysis *A high level examination of the concepts of the semantic Web and linked data is provided with a particular focus on the challenges they entail for libraries and their meta-data practices in the perspective of the extensive restructuring process that has already started.*
Conclusion *We demonstrate that a set of heterogeneity conflicts, threatening the level of semantic interoperability, can be associated with various phases of this restructuring process from analysis and modelling to conversion and external interlinking. It also claims that these conflicts and their potential solutions are mutually dependent across the phases.*

CHANGE FONT

Introduction

The report *On the record* (Library of Congress Working Group on the Future of Bibliographic Control, 2008) states that the '*library community's data carrier, MARC, is based on forty-year old techniques for data management and is out of step with programming styles of today*'. The report recommends future library standards to be integrated into a Web environment. Three years later Library of Congress followed up the conclusions from the report and announced that a '*new bibliographic framework project will be focused on the Web environment, Linked Data principles and mechanisms, and the Resource Description Framework (resource description framework)*' (Library of

Congress, 2011). In November 2012 the primer *Bibliographic Framework as a Web of Data: Linked Data Model and Supporting Services* (Library of Congress, 2012) was released providing an initial draft of a dedicated linked data model for bibliographic metadata (BIBFRAME in short).

Knowledge organizational approaches in the library community are increasingly characterized by a desire to harmonize with the Web architecture (Coyle, 2010; Hodge, 2000). During the last couple of years the community has developed a number of comprehensive metadata standardization projects inspired by the idea of linked data such as the BIBFRAME model. Linked data is a set of best practice principles of publishing and exposing data on the web utilizing a graph based data model powered with semantics and cross-domain relationships Ã¢â‚¬â€œ '*the semantic Web done right*' according to the web pioneer Sir Tim Berners-Lee (Heath, 2009).

In the light of traditional and current metadata practices of libraries the best practices of linked data imply a restructuring process from a collection of semi-structured bibliographic records to a semantic graph of unambiguously defined entities. Graphs are not new, neither as applied technology for knowledge organization (e.g. *The Network model*, a database model from the late 1960s) nor as a field of study (Graph theory as a mathematical field dates back Leonard Euler`s experimentations in the 1700s). Nevertheless, as a model for metadata structuring in libraries graphs introduce a new and challenging model for describing and organizing collections. This article demonstrates that the challenges can be associated with various phases of the restructuring process mentioned above - from analysis and modeling to conversion and external interlinking. Further it claims that these challenges and their potential solutions are mutually dependent across the phases. A poor initial analysis of the original model and the metadata that are designed according to this model could for example influence the design of a new (linked data model) and the final interlinking to external resources.

The concept of semantics is neither new in this context. HjÃƒÂ¸rland (2007) argues that semantic issues '*underlie all research questions*' in Library and Information Science and especially in the subfield Knowledge organization (KO). He also remarks that many consider the Semantic Web as one of the "important frontiers". The semantic Web is essentially an ambition to link data across different domains and to enable machines to act upon the links. The ambition requires that machines *understand* external data, or in other words that a minimum level of *semantic interoperability* is provided. Semantic interoperability is a key concept in this analysis of semantic Web orientated restructuring.

The article provides a high level examination of the concepts of the semantic Web and linked data, such as semantic interoperability. It has a particular focus on the challenges they entail for libraries and their metadata practices in the perspective of the extensive restructuring process that has already started. The examination is carried out through a review of relevant research within the field and of the essential documents that describe the key concepts.

The initial sections introduce and discuss the notions of graphs and semantic Web. The latter sections deal with the various phases of the restructuring process.

The giant global graph

Formally a graph G is a structure which consists of a set of nodes N and a set of edges E expressed as a pair, G = (N, E). The nodes represent objects, and the edges are relationships (or properties) connecting the nodes. An example of an applied graph is the World Wide Web which can

be regarded as a set of interlinked documents where each document is a node and the links are edges connecting the documents. This Web graph is used in Google`s PageRank algorithm to assign (relative) weighting to documents based on their incoming links (utilizing the Eigenvector centrality measure as described in Page, Brin, Motwani, and Winograd, 1999). Another example of a graph is a set of bibliographic metadata, where entities like authors, titles and year of publication is represented as nodes N = {Henrik Ibsen, A dolls`s house, 1879}, and the edges are properties relating the authors to the correct titles, and the titles to the year of publication E = {Henrik Ibsen-A doll`s house, A doll`s house-1879}.

After inventing the essential components of todayÃ¢â‚¬â„¢s Web architecture Sir Tim Berners-Lee later introduced the idea of an extension of the Web enabling not only relationships between documents but also between the things that the documents were about: In practice, a graph of interlinked data objects published and exposed on the Web. The idea was first presented as a *Semantic Web* (Berners-Lee, Hendler, and Lassila, 2001), then connected to a concrete technological infrastructure and a set of best practice publishing guidelines and revitalized as *linked data* (Berners-Lee, 2006). Berners-Lee has later used the terms Giant global graph (2007) and the *Web of data* (Bizer, Heath, and Berners-Lee, 2009) to express more or less the same concept. There are some discussion about the meaning of these terms, but a common interpretation is that the Semantic Web, the Giant Global Graph and the Web of data signify a high-level *vision*, whereas linked data represents the methods for realizing the vision (Heath, 2009).

Linked data

One of the main challenges in realizing a semantic Web is the heterogeneous nature of the metadata in various communities. An essential principle in the numerous guidelines for publishing linked data (Berners-Lee, 2006; Heath and Bizer, 2011; W3C, 2012) is therefore to use established standards like resource description framework. According to its suite of specifications [1] resource description framework provides a framework for representing resources as a set of statements based on a Graph data model. The statements consist of two nodes, a *subject* and an *object*, and a *predicate* that connects them. The statement '*Henrik Ibsen wrote A doll`s house*' can be outlined as a resource description framework statement where Henrik Ibsen is the subject, A doll`s house the object and the property wrote is the predicate. The three components form a *triple*, and a single resource description framework graph is the totality of such triples in a given universe of statements. There are some discussion on how well the resource description framework specifications are founded in the established mathematical concept of graphs (e.g. Hayes and Gutierrez, 2004), nevertheless the resource description framework graph is often characterized formally as a *directed labelled graph* since the edges always point from a subject towards an object and explicitly denote the property of the subjects. In order to make the resource description framework graphs machine processable and to integrate them with the Web architecture Uniform Resource Identifiers (URI) [2] are used to identify subjects, predicates and in some cases objects. Borrowing a predicate from the Dublin Core Metadata Terms [3] to label the edge in the example above, a triple based on URI`s can be expressed in triple notation as:

```
http://example.org/A_dolls_house http://purl.org/dc/terms/creator
http://example.org/henrik_ibsen .
```

Objects are also allowed to be literal values as "1979" in the following triple:

```
http://example.org/A_dolls_house http://purl.org/dc/terms/issued"1879" .
```

And the objects can be URIs created outside the local resource description framework graph as the URI from DBpedia [4] in the triple:

http://example.org/henrik_ibsen http://www.w3.org/2002/07/owl#sameAs http://dbpedia.org/resource/Henrik_Ibsen .

The three examples form a resource description framework graph as visualized in Figure 1.

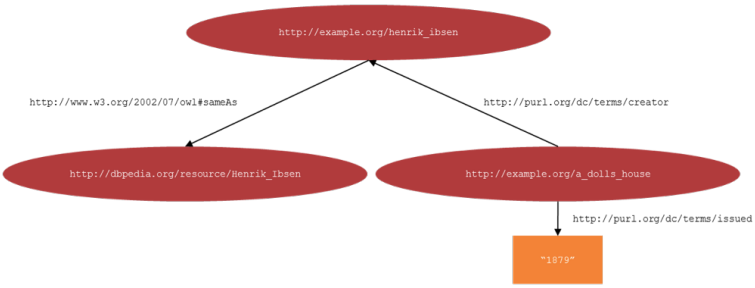

Figure 1 - A simple resource description framework graph of three triples

Interlinking of data

The basic resource description framework graph in Figure 1 also exemplifies some of the other essential principles of best practice linked data, using (HTTP) URIs as names for things being one of them [5]. In order to achieve a Giant global graph of truly interlinked data it is fundamental to provide links to URIs in external data sets. This is achieved in the example resource description framework graph by the link to a representation of the author Henrik Ibsen in the DBpedia data set which contains resource description framework structured information derived from Wikipedia. The property sameAs is taken from the Web Ontology Language (OWL) [6] and used as a predicate to denote the concurrence of the two representations of the author. In the vision of a Semantic Web such links based on HTTP URIs pointing to standardized data representations provides a platform for computational reasoning across institutions and communities. Reusing properties and classes from established and widely adopted vocabularies and ontologies, like Dublin Core and OWL, is considered a good practice which makes it easier to interpret and process the data for client applications. However linked data sources often mix self-defined and existing properties and classes.

2.3 Conceptualization

In the literature of linked data it is difficult to find a definite division between the terms vocabularies and ontologies. Gruber famously defined an ontology as a '*specification of a conceptualization*' (1993). The same broad definition could be used to describe a vocabulary (or a metadata schema). However ontologies tend to be used frequently to describe complex systems that provide a set of inference rules and description logic enabling computational reasoning, while vocabularies are used quite consistently to describe less complex collections of conceptual terms like the aforementioned Dublin Core Metadata Terms. In the field of Knowledge representation TBox and ABox are often used to separate between a high level representation system and the actual instance data generated in accordance with such systems (Bergman, 2009; Ferrara, Lorusso, Montanelli, and Varese, 2008). TBox (T for *Terminological*) constitutes a set of concepts, properties and constraints on their usage. ABox (A for *Assertions*) constitutes a set of assertions which are structured according to the TBox, for example, a collection of resource description framework-triples. Within a slightly broad definition, both ontologies, vocabularies, metadata schemas and standards providing some sort of concepts, properties and constraints could be defined as a TBox.

Semantic interoperability

At the core semantic Web and linked data are about connecting data across heterogeneous domains enabling computers to understand the data and their relations. '*[Ã¢â‚¬Â¦] Information is given well-defined meaning*' (Berners-Lee, Hendler, and Lassila, 2001) and this meaning is enhanced with machine-interpretability by the use of standards like resource description framework, unique identifiers and referenced ontologies (as described above in section 2).

Some have questioned such a definition of semantics (Uschold, 2003), and others have discussed whether itÃ¢â‚¬â„¢s in accordance with established approaches in computer science and linguistics (Almeida, Souza, and Fonseca, 2011; Sheth, Ramakrishnan, and Thomas, 2005). Regardless of these objections and discussions, it is natural to associate an operational understanding of semantics in the context of linked data with the overall goal to provide *interoperability* across heterogeneous domains.

While interoperability in general can be defined as the ability of two or more systems to exchange information and to use this information, *semantic interoperability* specifies a certain requirement to achieve this goal: The ability of two or more systems to exchange and share intended meaning (Kalfoglou, 2010; Nilsson, 2010; Park, 2006). Semantic interoperability often constitutes one level in a conceptual model which distinguishes it from other levels of interoperability such as syntactic interoperability concerning exchange formats and technical interoperability concerning exchange protocols (see e.g. Nilsson, Baker, and Johnston, 2009; Tolk and Muguira, 2003; Tolk, 2006).

For a system to understand the intended meaning of information in other systems, the information being exchanged needs to be equipped with a minimum of disambiguous machine-interpretable description. In a linked data conformant resource description framework graph the interpretable description is to be found in the referenced ontologies defining the meaning of certain properties and classes, as described and exemplified above in Figure 1. GarcÃƒÂa-Castro and GÃƒÂ³mez-PÃƒÂ©rez (2011) provide a definition of semantic interoperability where this functionality is outlined explicitly as '*the ability that semantic systems have to interchange ontologies and use them*'.

The challenges to achieve semantic interoperability can also be defined and explained negatively by the existence of a certain degree of *semantic heterogeneity* between two systems. Pluempitiwiriyawej and Hammer (2000) have classified occurences of semantic heterogeneities in XML data sources. Some of their main conflict classes can be related to conflicts arising in the process of interlinking instances described with disparate ontologies:

- *structural conflicts*, when the same (or overlapping) classes or properties are represented differently in two ontologies due to discrepancies in the level of generalization/specialization
- *data conflicts*, when the same concept is represented differently due to incorrect spelling and different identification systems

Ferrara, Lorusso, Montanelli and Varese (2008) highlights three sources of heterogeneity challenging the matching of instances across populated ontologies: structural heterogeneity, data value differences, and logical heterogeneity. The first two equals the structural conflicts and data conflicts mentioned above. The latter is concerning differences in the way ontologies are implementing rules for reasoning. In addition to these conflicts Ferrara (2005) has described semantic heterogeneity scenarios related to flexible schemas providing semi-structured data, where

conflicts arises from the inconsistencies in usage and interpretation of the schema rules.

Semantic heterogeneity conflicts are potential obstacles to achieving the degree of semantic interoperability necessary for a successful realization of the Semantic web. Bizer, Heath, and Berners-Lee (2009) have announced data fusion and schema mapping to be one of the main research challenges related to linked data. The next section discusses the potential obstacles in relation to the restructuring of library data.

The linked data restructuring cycle

Cultural heritage institutions like libraries possess huge amounts of metadata already catalogued and stored according to the principles of established community standards. Representing these data as resource description framework graphs and linking them '*to other people `s data*' (Berners-Lee, 2006) leads into a cycle of restructuring. This cycle can be derived from the best practice guidelines and is analysed and described in detail by the LOD2 project [7] and by other parties (e.g. Hyland, 2010; W3C, 2012). In the context of a concrete case study of restructuring library data TallerÃƒÂ¥s, Massey, Dahl, and Pharo (2013) have synthesised (and simplified) existing efforts in describing a linked data restructuring cycle (see Figure 2). With an exception of the evaluation aspects each of the phases in the cycle will be discussed in separate subsections, with a special focus on the case of restructuring library data. The cycle can be viewed as an iterative process with the starting point in an analysis of a certain domain. A new ontology is developed; the data is converted in accordance with this ontology and interlinked with data in other datasets published on the web. The latter phase can be considered as an on-going evaluation with the potential to restart the process initiating a deeper analysis, remodelling of the ontology and a tuning of the conversion algorithm and interlinking technique.

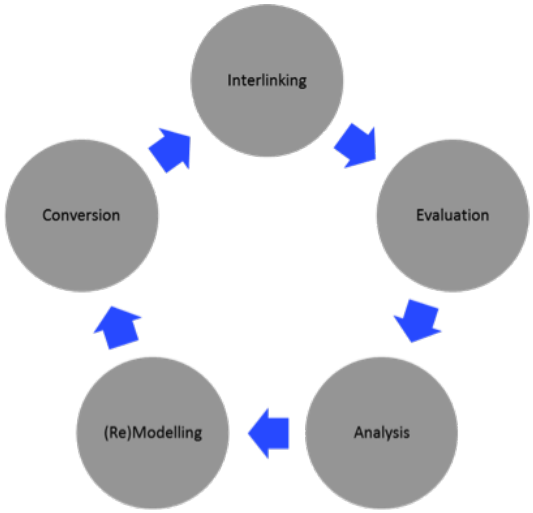

Figure 2 - Linked data restructuring cycle

Analysis: Library data

Parallel to the developments at the Library of Congress, described in the introduction, the library community has witnessed a great number of '*bottom up*' linked data initiatives. The national libraries of Sweden (Malmsten, 2009), France (BibliothÃƒÂ¨que nationale de France, 2012), Germany (Hauser, 2012) and Great Britain (The British Library, 2013) have all carried out major projects involving a conversion of their catalogue data into a variety of resource description framework implementations. OCLC have made a data set of the three top levels of the Dewey Decimal Classification System in 23 languages available as linked data [8]. They also host the Virtual International Authority File (VIAF) project connecting authority records from several national and

other libraries, also made available as resource description framework[9]. See Dunsire and Willer (2011) for an extensive overview of other linked data projects in the library community.

These projects reveal a desire for change, and a belief in the vision of the Giant global graph. What are then the dissatisfactory aspects of the existing traditions of metadata production motivating such desires and beliefs? And what challenges concerning restructuring are to be found in traditional library metadata?

The bibliographic record

Since the middle of the 1800s *universal bibliographic control* (universal bibliographic control) has been an expressed objective in the library community (Svenonius, 2000). universal bibliographic control is the vision of a shared worldwide bibliography of every book ever published. To support this vision library history has offered different bibliographic systems based on available technology. These systems have undergone two major revolutions, respectively, the transition from the book catalog to the card catalog and from the card catalog to the automated systems that characterize current practices. The first revolution was the origin of the independent bibliographic record in form of a card containing description of a certain edition of a book. The second revolution automated this record and made it '*machine readable*' (Avram, 1975). The struggle to achieve universal bibliographic control has emphasized standards in order to support interoperability and exchange of bibliographic records between the contributing libraries Ã¢â‚¬â€œ the ideal has been to catalogue a book only once. The standards have also changed in accordance with the bibliographic systems they were developed to support. Today the most widespread standards are the cataloging rule *Anglo-American Cataloging Rule* (AACR) [10] and the metadata schema Machine Readable Cataloguing (MARC) [11]. Both standards were developed during the 1960s.

These standards have increasingly been criticized for several reasons. The general critique concerns their age and that they are out of step with the '*programming styles of today*' (as stated in the report issued by Library of congress cited in section 1). The standards were developed prior to relational databases (Codd, 1970), and the Web, and lack important knowledge organizational innovation from those infrastructures; for instance the idea of using unique and computable identifiers like database keys and URIs. Instead they are tightly intervened with some of the knowledge organizational principles implied in the card catalogue, the leading technology of the time they were developed (Coyle and Hillmann, 2007; Thomale, 2010). This includes carrying on the principle of bibliographic records being geared for human reading and interpretation, resulting in semi-structured MARC records containing mostly text strings. These strings are machine readable, but harder to reason upon for machines than well-structured data in accordance with relational database theory or description logic in ontologies (Styles, Ayers, and Shabir, 2008). The restructuring phases described below all concerns successful identification of entities. Inconsistent cataloging due to heterogeneity conflicts in terms of data values and human interpretation of standards may lead to both data loss, where the text is not understood by the machine, and redundancy, where two or more text strings in a given data set are representing the same entity (described in more detail in section 4.3).

The bibliographic MARC record also continues the principle of describing a certain edition of a book (a *manifestation* of a work in terms of the FRBR model (IFLA Study Group on the Functional Requirements for Bibliographic Records, 1998)). The lack of consistent references to the platonic idea of a work entity that connects the potential manifold of

editions, and the lack of unique identifiers of authors, publisher etc., constitutes a data model of rather unconnected and segregated records ('*islands of data*'), representing the opposite of the idea of a unified data set based on directed and (semantic) labeled graph connecting data objects.

Some have faced this criticism and argue that many of the problems are to be found in the lack of complexity within the systems that manage the MARC records, for example in utilizing the sophistication of relationships expressed in the card catalog (e.g. Murray and Tillett, 2012). Library of Congress and other major stakeholders have nevertheless, as outlined above, regarded the problems to be too extensive to make mere adjustments.

(Re)modelling library data: universal bibliographic control vs. the giant global graph

The primary method to achieve universal bibliographic control has been standardization. Groups of experts from leading institutions like the Library of Congress have developed and maintained the standards, and consistency has been secured by the principle of everyone using these standards. In this perspective universal bibliographic control can be described as a '*top down*' approach to interoperability. Linked Data represent a more pragmatic and '*bottom-up*' approach. Berners-Lee, Hendler, and Lassila (2001) states that the Semantic Web '*will be as decentralized as possible*'. When the new ontologies within different domains and communities are to be designed, metadata managers are free to choose and mix classes and predicates from existing ontologies and vocabularies with their own terms.

Broad definitions of TBox and ABox (as given in Section 2.3) may be useful in a comparative analysis of e.g. OWL-based ontologies and domain specific standards from the library world. Ontologies should according to guidelines for Linked Data and Semantic Web facilitate automated reasoning. This requires that the ontologies describe concepts, properties and rules for their usage in a processable way. Library standards do indeed provide concepts, properties and rules for using them, but they tend to - especially the rules (such as AACR2) - to be oriented towards human consumption and not automated reasoning. To make a good linked data model, it is important that it not only ensures a successful conversion of the instance data in the ABox, but also a machine readable TBox.

Svenonius (2000) has remarked that the bibliographic records were made to support a fixed set of functions and objectives, such as inventory and the objectives formulated by prominent catalogue innovators like Cutter and Lubetzky, and that technological advance and new media formats have challenged these functions: '*It is hard surprising that using one device to serve several functions should lead to trouble in times of technological change*'. In Rust and Bide (2000) such conflicts related to intended function of the data (retrieval aspects, cataloguing aspects etc.) is outlined as potential interoperability obstacles.

Through a bottom up approach libraries get the opportunity to handle such obstacles. Different ontologies can be designed according to the needs at the time they occur. This possibility is also utilized in the different linked data projects described above. If the community however wishes to maintain the idea of universal bibliographic control, as in a worldwide bibliography based on distributed contributions, they also need to balance this flexibility with the actual potential for operational interoperability between the ontologies in use [12]. This also involves technical challenges related the long term archiving of resource description framework data as discussed in Seadle (2013). Lately there has been a lively and interesting discussion about '*Reuse (or not) of existing ontologies*' at the BIBFRAME email list, where supporters of a

flexible '*bottom up*' approach arguing for reuse opposites supporters of a new and local model arguing for control and long term sustainability ('*BIBFRAME archives*,' 2013).

Conversion

Case studies of mapping library records to resource description framework based ontologies have confirmed all of the potential semantic heterogeneity conflicts mentioned in section 3.1., such as inconsistencies and structural discrepancies (TallerÃƒÂ¥s, Massey, Dahl, and Pharo, 2013; Westrum, Rekkavik, and TallerÃƒÂ¥s, 2012). They have also shown that such conflicts have a serious impact on the conversion of data. Without unique identifiers for the various entities the conversion is dependent on a computational interpretation of the strings. The entities, for instance represented by the string Ibsen, Henrik from the field for main entries in the MARC record, are reduced to a set of characters to be matched with other sets of characters. Then a decision is made, based on a chosen similarity threshold, as to whether the characters represent the same entity or not. If the similarity measure satisfies the threshold a URI can be assigned as a unique and single identifier for this entity. The URI is further assigned into a series of triples of the kind exemplified in section 2.1. All forms of inconsistencies due to misspellings, cultural or linguistic contexts or different interpretations of the rules on how to describe things, affect such a process negatively, and will make the conversion algorithm fail to assign correct identifiers.

To improve the result of the conversion process some argue that one should use ontologies based on terms exclusively from local schemas, such as a MARC based ontology[13], in order to overcome structural heterogeneity, secure semantic coherence and reduce the potential *lossyness* in the conversion process (Dunsire, 2012). This argument is more in line with the traditional *top down* universal bibliographic control approach emphasizing domain specific standards. Others have worked with pre-coordination of existing MARC collections in order to harmonize them to other prominent data models in the community such as the FRBR model, and in order to *clean* the records and reduce inconsistencies prior to the conversation (and interlinking) process (Aalberg, 2006; Westrum, Rekkavik, TallerÃƒÂ¥s, 2012). Nilsson (2010) have described the latter approach as a vertical harmonization within a certain domain.

Interlinking

When the data is converted to an resource description framework format, they should be linked to similar data in existing resource description framework graphs. Many of the data sets that already have been published as Linked Data describe cultural objects and entities related to them. These data sets are largely overlapping with library data, which constitutes a great potential for an extensive interlinking. The main challenge in this part of the restructuring process is once again related to semantic interoperability and the question of how to decide which URIs that are representing the same concept or the same entity in two different data sets that are structured according to different ontologies. Although this is an area under development, there are already a number of automated methods for approaching the problem. They range from simple string recognition techniques (often referred to as naÃƒÂ¨ve interlinking) to utilizing the graph structures in the resource description framework graphs, machine learning techniques and more probabilistic oriented methods (for some examples see Doan, Madhavan, Dhamankar, Domingos, and Halevy, 2003; Melnik, Garcia-Molina, and Rahm, 2002; Raimond, Sutton, and Sandler, 2008). In practice the interlinking of resource description framework graphs is a semi-automatic discovery phase, both dependent on manual and automatic approaches. The manual efforts can be related to supervision of automatic systems, but also to

direct assignments of links, for instance in the cataloguing process.

Similarity and identity are not fixed categories, albeit the extensive use of the rather unambiguous predicate owl:sameAs to express concurrences in the Linked Data context (defined in OWL as: '*an statement [that] indicates that two URI references actually refer to the same thing*'][14]). Halpin, Hayes, McCusker, McGuinness, and Thompson (2010) claims that linked data experience an '*identity crisis*': '*Just because a construct in a knowledge representation language is explicitly and formally defined does not necessarily mean that people will follow that definition when actually using that construct Ã¢â‚¬Ëœin the wildÃ¢â‚¬â„¢*'. Based on a logical perspective on identity ('*Leibnitz`s law*') they identify a variety of inconsistent usage cases of the owl:sameAs predicate and a number of reasons for them. Some of these can be related directly to heterogeneity conflicts such as discrepancies in the interpretation of flexible ontologies. Bizer, Heath, and Berners-Lee (2009) also address the problem of structural heterogeneity claiming that existing correspondences often are too '*coarse-grained*' to support effective computational reasoning.

Concluding remarks

A proper analysis of existing data, the standards used to generate them and the domain specific needs and objectives forms the basis for the development of a new data model. This data model must maintain the basic semantics from the existing standards, and at the same time aim to innovate and renew old traditions. The quality of the conversion from the old to the new model depends on how well the model is able to handle heterogeneity conflicts in order to maintain granularity and semantic attributes, and eventually prevent significant loss of data (and semantics). The semantic expressiveness in the new model is also vital for providing precise links to other dataset.

Through references to research, standards and best practice-documents the article have outlined a restructuring process from a record-based data model to best practice linked data. Library data is used as a case to discuss challenges in the various phases of the process. Library data is an interesting case because the library community is already in an active process of restructuring. Each of the phases represents specific challenges regarding semantic heterogeneity conflicts, but these challenges also connect the phases and make them mutual dependent. The quality of the implementation of each phase will influence on the ability to gain quality in the other phases.

In a future research project it would be interesting to conduct a more thorough examination of concepts such as semantic interoperability and heterogeneity conflicts. In the research literature there exist a manifold of definitions and interpretations, other than those outlined in this article. A classification of these definitions, based on context and specific technological challenges, could for instance be useful in order to establish a fruitful theoretical perspective on the semantic Web.

[1] In particular resource description framework Primer (http://www.w3.org/TR/rdf-mt/), resource description framework Concepts and Abstract Syntax (http://www.w3.org/TR/rdf-concepts/) and resource description framework Semantics (http://www.w3.org/TR/rdf-mt/)

[2] http://www.w3.org/TR/rdf-primer/#identifiers

[3] http://dublincore.org/documents/dcmi-terms/

[4] http://dbpedia.org/About

[5] To gain a seamless Web integration the guidelines recommend HTTP based

URIs.

[6] http://www.w3.org/TR/owl-ref/

[7] http://stack.lod2.eu/

[8] http://dewey.info/

[9] http://viaf.org

[10] http://www.aacr2.org/

[11] http://www.loc.gov/marc/

[12] Lately there has been a lively and interesting discussion about Reuse (or not) of existing ontologies' at the BIBFRAME email list: http://listserv.loc.gov/cgi-bin/wa?A1=ind1303&L=bibframe

[13] See http://marc21rdf.info/ for a resource description framework based Vocabulary representing MARC elements

[14] http://www.w3.org/TR/owl-ref/#sameAs-def

References

Aalberg, T. (2006). A Tool for Converting from MARC to FRBR. *ERCIM News*

Almeida, M., Souza, R., and Fonseca, F. (2011). Semantics in the Semantic Web: a critical evaluation. *Knowledge Organization*, **38**(3), 187Ã¢â‚¬Å“203

Avram, H. D. (1975). MARC, its history and implications. *Washington, DC: Library of Congress*

Bergman, M. (2009). The Fundamental Importance of Keeping an ABox and TBox Split. AI3. Retrieved from http://www.mkbergman.com/489/ontology-best-practices-for-data-driven-applications-part-2/

Berners-Lee, T. (2006). Linked data: design issues. W3C

Berners-Lee, T. (2007). Giant Global Graph. Decentralized Information Group Breadcrumbs. Retrieved from http://dig.csail.mit.edu/breadcrumbs/node/215

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). *The Semantic Web. Scientific American*, **284**(5), 34Ã¢â‚¬Å“43

BIBFRAME archives. (2013). Retrieved from http://listserv.loc.gov/cgi-bin/wa?A1=ind1303&amP;L=bibframe

BibliothÃƒÂ¨que nationale de France. (2012). data.bnf.fr. Retrieved from http://data.bnf.fr/

Bizer, C., Heath, T. & Berners-Lee, T. (2009). Linked Data: The story so far. *International Journal on Semantic Web and Information Systems*, **5**(3), 1Ã¢â‚¬Å“22

The British Library. (2013). Free Data Services. The British Library. Retrieved March 1, 2013, from http://www.bl.uk/bibliographic/datafree.html

Codd, E. F. (1970). A relational model of data for large shared data banks. *i*, **13**(6), 377Ã¢â‚¬Å“387

Coyle, K. (2010). Library Data in a Modern Context. *Library Technology Reports*, **46**(1), 5Ã¢â‚¬Å“13

Coyle, K. & Hillmann, D. (2007). Resource Description and Access (RDA). *D-Lib Magazine*, **13**(1/2)

Doan, A., Madhavan, J., Dhamankar, R., Domingos, P., smd Halevy, A. Y. (2003). Learning to match ontologies on the Semantic Web. *The VLDB Journal*, **12**(4)

Dunsire, G. (2012). An introduction to open linked data for librarians. Powerpoint presentation at the National Library of Finland

Dunsire, G. & Willer, M. (2011). Standard library metadata models and structures for the Semantic Web. *Library Hi Tech News*, **28**(3), 1Ã¢â‚¬Å“12

Ferrara, A., Lorusso, D., Montanelli, S. & Varese, G. (2008). Towards a benchmark for instance matching. *Proceedings of the 3rd International Workshop on Ontology Matching* (OM-2008)

GarcÃƒÂa-Castro, R.& GÃƒÂ³mez-PÃƒÂ©rez, A. (2011). Perspectives in semantic interoperability. *Proceedings of the International Workshop on Semantic Interoperability IWSI 2011 In conjunction with ICAART 2011* (pp. 13Ã¢â‚¬Å“22). SciTePress

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, **5**(2), 199Ã¢â‚¬â€œ220
Halevy, A. (2005). Why Your Data WonÃ¢â‚¬â„¢t Mix - New tools and techniques can help ease the pain of reconciling schemas. *ACM Queue*, **3**(8)
Halpin, H., Hayes, P., McCusker, J., McGuinness, D. & Thompson, H. (2010). When owl:sameAs isnÃ¢â‚¬â„¢t the same: An analysis of identity in Linked Data. *The Semantic Web Ã¢â‚¬â€œ ISWC 2010* (Vol. 6496, pp. 305Ã¢â‚¬â€œ320). Berlin: Springer
Hauser, J. (2012). Dokumentation des Linked Data Services der DNB. Retrieved from https://wiki.dnb.de/display/LDS/Dokumentation+des+Linked+Data+Services+der+DNB
Hayes, J. & Gutierrez, C. (2004). Bipartite graphs as intermediate model for resource description framework. *The Semantic Web Ã¢â‚¬â€œ ISWC 2004* (Vol. 3298, pp. 47Ã¢â‚¬â€œ61). Berlin: Springer
Heath, T. (2009). Linked Data? Web of Data? Semantic Web? WTF? Tom HeathÃ¢â‚¬â„¢s Displacement Activities. Retrieved from http://tomheath.com/blog/2009/03/linked-data-web-of-data-semantic-web-wtf/
Heath, T. & Bizer, C. (2011). Linked Data: Evolving the Web into a global data space. *Morgan & Claypool*
HjÃƒÂ¸rland, B. (2007). Semantics and knowledge organization. *Annual review of information science and technology*, **41**(1), 367Ã¢â‚¬â€œ405
Hodge, G. (2000). Systems of knowledge organization for digital libraries: Beyond traditional authority files. *Washington: The Digital Library Federation Council on Library and Information Resources*
Hyland, B. (2010). Preparing for a linked data enterprise. Linking enterprise data (pp. 51Ã¢â‚¬â€œ64). *Springer US*
IFLA Study Group on the Functional Requirements for Bibliographic Records. (1998). Functional requirements for bibliographic records: Final report. *MÃƒÂ¼nchen: K.G. Saur*
Kalfoglou, Y. (Ed.). (2010). Cases on semantic interoperability for information systems integration: Practices and applications. *New York: Information science reference*
Page, L., Brin, S., Motwani, R. & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the Web. *Stanford InfoLab*
Library of Congress. (2011). A Bibliographic Framework for the digital age. Retrieved from http://www.loc.gov/marc/transition/news/framework-103111.html#ftn1
Library of Congress. (2012). Bibliographic Framework as a Web of data: linked data model and supporting services. *Washington DC*
Library of Congress Working Group on the Future of Bibliographic Control. (2008). On the record: Report of the Library of Congress Working Group on the Future of Bibliographic Control
Malmsten, M. (2009). Exposing library data as linked data. *IFLA satellite preconference*
Melnik, S., Garcia-Molina, H. & Rahm, E. (2002). Similarity flooding: A versatile graph matching algorithm and its application to schema matching. *Proceedings of the 18th International Conference on Data Engineering*
Murray, R. J. & Tillett, B. B. (2012). Cataloging theory in search of graph theory and other ivory towers. *Information Technology and Libraries*, **30**(4), 170Ã¢â‚¬â€œ184
Nilsson, M. (2010). From interoperability to harmonization in metadata standardization - Designing an evolvable framework for metadata harmonization. *Royal Institute of Technology, Stockholm*
Nilsson, M., Baker, T.& Johnston, P. (2009). Interoperability levels for Dublin Core Metadata. Dublin Core Metadata Initiative. Retrieved from http://dublincore.org/documents/interoperability-levels/
Park, T. (2006). Semantic interoperability and metadata quality: An analysis of metadata item records of digital image collections. *Knowledge organization*, **33**(1), 20Ã¢â‚¬â€œ34
Pluempitiwiriyawej, C.& Hammer, J. (2000). A classification scheme for semantic and schematic heterogeneities in XML data sources. *Technical report TR00-004*
Raimond, Y., Sutton, C.& Sandler, M. (2008). Automatic interlinking of music datasets on the semantic web. *Linked Data on the Web - LDOW 2008*
Rust, G., & Bide, M. (2000). The indecs metadata framework: Principles, model and data dictionary. *Indecs Framework*
Seadle, M. (2013). Archiving in the networked world: resource description framework. *Library Hi Tech*, **31**(1), 182Ã¢â‚¬â€œ188
Sheth, A., Ramakrishnan, C. & Thomas, C. (2005). Semantics for the semantic web: The implicit, the formal and the powerful. *International journal on Semantic Web and*

information systems, **1**(1)
Styles, R., Ayers, D. & Shabir, N. (2008). Semantic MARC, MARC21 and the Semantic Web. *Linked Data on the Web - LDOW 2008*
Svenonius, E. (2000). The Intellectual Foundation of Information Organization. *Cambridge, Massachusetts: The MIT Press*
Tallerås, K., Massey, D., Dahl, J.H.B. & Pharo, N. (2013). Ordo ad chaos - Linking Norwegian black metal. Libraries, black metal and corporate finance: Current research in Nordic Library and Information Science (pp. 136Ã¢â‚¬â€œ150). *Borås: Univeristy of Borås*
Thomale, J. (2010). Interpreting MARC: WhereÃ¢â‚¬â„¢s the bibliographic data? *Code4lib*, (11)
Tolk, A. (2006). What comes after the Semantic Web - PADS implications for the dynamic Web. *PADS*, 55Ã¢â‚¬â€œ62
Tolk, A. & Muguira, J. (2003). The levels of conceptual interoperability model. *Proceedings of the 2009 Spring Simulation Multiconference*
Uschold, M. (2003). Where are the Semantics in the Semantic Web? *AI Magazine*, **24**(3), 25Ã¢â‚¬â€œ36
W3C. (2012). Linked data cookbook. Retrieved from http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook
Westrum, A., Rekkavik, A. & Tallerås, K. (2012). Improving the presentation of library data using FRBR and linked data. *Code4Lib Journal*, (16)

How to cite this paper

Tallerås, K. (2013). From many records to one graph: heterogeneity conflicts in the linked data restructuring cycle. *Information Research*, **18**(3) paper C18. [Available at http://InformationR.net/ir/18-3/colis/paperC18.html]