Marta Cichoń

# LIBRARY DATA VISIBILITY AND RE-USE: POSSIBILITIES EMERGING FROM THE NATIONAL LIBRARY DESCRIPTORS PROJECT

## Introduction

As the competitive advantage that can be achieved by increasing data visibility through services like Google or Bing gets more and more recognised, libraries – as service and data providers – are facing challenges similar to those of other organisations that want to benefit from big data trends. Within these trends, broadening the access to data and enabling the integration of information from multiple data sources are commonly understood necessities. Since researchers nowadays usually want to take advantage of the possibilities provided by aggregated resources, which in turn allow them to obtain the productivity (and often research results) in an incomparably more efficient way, the requirement to organise library data as semantically related datasets becomes more and more imperative. Achieving this goal requires aggregating and combining data from different sources. The problem, which many libraries are exposed to, is that the integration of their data is prevented by legacy systems and incompatible standards and formats. The National Library of Poland currently stores its data – like many other libraries – in proprietary formats such us MARC 21 understood generally by the library community and nobody beside it, which makes this data not easily reusable with other data stores accessible through the Web.

The National Library Descriptors project launched at the "National Library Descriptors" Conference on April 20–21, 2015, aims first and

foremost at providing better data access by creating additional access points within the original National Library dataset of the Bibliographic Database. This aim is to be achieved by providing better data granularity and segmentation within this dataset stored in the MARC 21 Format by using some additional MARC 21 properties – as-yet unused (or even unavailable within the format) – in bibliographic records, such as "Audience Characteristics" (MARC 21 field 385), "Creator/Contributor Characteristics" (MARC 21 field 386) or "Time Period of Creation" (MARC 21 field 388). In the same way, in order to create additional access points to the authority data, a set of additional attributes has been defined based on the additional MARC 21 fields added to the MARC 21 Format by the Library of Congress, such as "Associated Place" (MARC 21 field 370), "Field of Activity" (MARC 21 field 372), "Associated Group" (MARC 21 field 373), "Occupation" (MARC 21 field 374), or "Gender" (MARC 21 field 375).

Looking ahead, the project implies not only providing and maintaining these additional access points within a library catalogue, but also using the newly added attributes as additional relations – previously unavailable due to the lack of authority controlled access points – between various entities stored within the database (such as Personal Names, Organisations, Geographic Names and Publications). Through simultaneous efforts toward developing the Linked Data model corresponding with the National Library's set of bibliographic data and publishing this data as the open RDF dataset, it is currently regarded that these additional relations could in the future provide further advantages in combining the National Library dataset with other datasets available on the Web, as they can be relatively easily mapped to properties and classes defined in the commonly used Semantic Web ontologies, in comparison to the data expressed in the information retrieval language currently used by the National Library. They may also provide additional contexts for possible NER tool implementation that could be applied to the digitised content of the National Digital Library. For the very same reason, the National Library Descriptors project implies simplification of the controlled vocabulary used as the subject headings in the bibliographic database. It is understood that the transition to an effectively integrated dataset requires accessible data structures and modelling of the data within the RDF schema in order to enable interlinking with other data stores. The National Library Descriptors project is expected to provide a gradual approach to facilitate that result.

## Current data trends

Management of information and knowledge has been transformed in recent decades. In addition to the shift towards digital management of information, movements adopting and advocating open approaches to share these digital resources have been emerging. At this point it seems commonly understood that all groups of information consumers are interested in the instrumental value of open access to data, but open data is just one aspect of the "data revolutions" taking place nowadays.[1] Currently, Internet datasets are created on a large scale, often in a non-standard way. Additionally, the development of IT infrastructure allows for their arbitrary expansion. Most of the data collected in this way constitute open and widely available datasets, which become priceless sources of information for further processing and analysis.[2] Precise processing of large datasets is considered a major challenge in information management and data analysis, as well as in related areas. Throughout the greater part of human history, only small portions of data could be analysed due to lack of appropriate tools that would allow the acquisition, organisation, storage and processing of information effectively. Despite an enormous shift in the approach to data collection and distribution, there are still lots of legacy practices in this field, resulting from past practices and institutional structures that assume that access to information is limited.[3]

While access to data still needs to be broadened, achieving immense benefits from this information requires combining data from different sources – often from organisations that have no history of sharing data at scale. In the era of "Big Data" the challenge lies not only in processing large-scale sources of data, but also in shifting the paradigm of data collection, taking into consideration the possibility of acquiring and integrating multiple types of data from various, often very different sources. Connecting diverse sources of data in order to achieve synergy effects leading to the production of new information and, consequently, new knowledge, is already a recognised challenge for the immediate future. Linking "traditional" data sources, such as public and research data, with

1 T. Davies, D. Edwards, 'Emerging Implications of Open and Linked Data for Knowledge Sharing in Development', *IDS Bulletin*, vol. 43 (5), 2012, pp. 117–127, http://dx.doi.org/10.1111/j.1759-5436.2012.00372.x [access: 23/03/2016].

2 *Internet: publiczne bazy danych i Big data* [Internet: public databases and Big data], ed. G. Szpor, Warszawa 2014, p. 52.

3 V. Mayer-Schönberger, K. Cukier, *Big data: rewolucja, która zmieni nasze myślenie, pracę i życie* [Big data: a revolution that will transform how we live, work and think], Warszawa 2014, p. 36.

new sources of data, such as various web services, might be a unique opportunity for complex exploration of social and cultural behaviours and newly emerging phenomena. While the benefits of such synergy are most present in the social sciences, they are also definitely feasible in most other areas of research. Nevertheless, to take full advantage of the prospectively linked data sources, some difficulties still need to be overcome.[4] Still, although we continue to be constrained by limited resources, profiting from all the available data is already reasonable and viable in a growing number of cases in domains where it has not been feasible so far. Often, the latent value of information might be discovered only by linking one dataset with another one, as already mentioned, even when they seem entirely divergent at first sight. Such an approach allows for the creation of innovative solutions based on blending data in novel ways.[5]

Researchers are nowadays overwhelmed by vast amounts of information. This information can come from many distributed sources, and in many cases it is far beyond what we can deal with on our own. As a result, there is an increasing demand for automated and semi-automated systems that sort through and assimilate this informational excess, allowing for further re-use by machines or people. The high-level goal is to create an assimilator that would act as an intermediary between humans and information. The assimilator would get queries from a human and then gather information from all relevant sources by culling through it as accurately as possible. Such a system would combine all that bears usefully on what the human wants to know, and provide the human with a coherent solution that corresponds to their intent.[6] Currently available technology allows for accessing databases through specialised Web interfaces. Still, more and more researchers want to use collections as a whole, mining and organising the information in alternative ways.[7]

In the present world of data, the sum of information is more valuable than any of its parts, and the same rule may be applied to linked datasets. Nowadays, Internet users are already familiar with mashup web services, which are basically web sites presenting information from at least two or

4 *Internet: publiczne bazy danych i Big data*, op. cit., p. 55.

5 V. Mayer-Schönberger, K. Cukier, op. cit., pp. 55, 144.

6 H. Haidarian Shahri, *On the Foundations of Data Interoperability and Semantic Search on the Web*, dissertation, University of Maryland, 2011, p. 9, http://hdl.handle.net/1903/11798 [access: 23/03/2016].

7 L. Johnston, 'Digital Collections as Big Data', Digital Preservation Meeting, Library of Congress, 2012, www.digitalpreservation.gov/meetings/documents/ndiipp12/BigData_Johnston_DP12.pdf [access: 23/03/2016].

more sources in an innovative way, often displaying data in a visual form, making it even more accessible to the broader public. One of the methods of data re-use is rearranging and restructuring it as if designing it from scratch in order to enable data extension, which makes it suitable for further and multi-way re-use.[8] This approach has been adopted during the design phase of the National Library Descriptors project. How exactly it might influence the re-use of the corresponding data will be explained further.

As modern science continues its exponential growth in complexity and scope, the need for more collaboration among scientists at different institutions in various subareas and across scientific disciplines is becoming increasingly important. Researchers working at one level of analysis may need to find and explore results from another level, another part of the field, or from a completely different scientific area.[9] One of the difficulties yet to be overcome for providing better access to these results is legacy systems with incompatible standards and formats, which often prevent the integration of data. The implementation of successive technologies over decades has scattered the metadata of libraries, archives and museums across multiple databases, spreadsheets, and even unstructured word processing documents. For business continuity reasons, legacy and newly introduced technologies often coexist in parallel. Even in cases where a superseded technology is completely abandoned, relics of the former tool can often be found in the content, which has been migrated to the new application.[10] It has been a truth generally acknowledged, and for quite a long time, that scientists are becoming increasingly reliant on Internet resources for supporting their research endeavours. In the search for a domain-specific Web site or a paper on a particular topic, web-engines can do a phenomenal job of sorting through billions of possibilities and identifying potentially useful search results. As a result, the Web has become indispensable for supporting traditional communication within various knowledge disciplines as well as serving the needs of scientists within their own disciplinary boundaries.[11] However, there is still the "Invisible Web", otherwise known as "Deep Web", a term coined for searchable databases that consist of the component parts of the Web which are

8 V. Mayer-Schönberger, K. Cukier, op. cit., p. 146.

9 J. Hendler, 'Science and the Semantic Web', *Science*, 24/01/2013, vol. 299, issue 5606, pp. 520–521, http://science.sciencemag.org/content/299/5606/520.full [access: 23/03/2016].

10 S. van Hooland, R. Verborgh, *Linked Data for Libraries, Archives and Museum: How to clean, link and publish your metadata*, London 2014, pp. 1–2.

11 J. Hendler, op. cit., p. 520.

hard or impossible to find through prominent search engines and directories. While these items remain outside the scope of traditional Web search tools, they still reside "on" the Web. Generally, the Invisible Web is comprised of records in databases, with most library databases among these. The dramatic term "invisible" is meant to underscore the importance of realising that there is more to the Web than a Meta-crawler search might reveal.[12] The more these compartments of the Web are linked to existing open datasets in order to facilitate access to these compartments, the less invisible the resources become, and the richer and more productive our experience using the Web for research. In order to link data across the Web we need to be able to interconnect data across independent islands. The word "islands" is used here to emphasise that each information system is modelled for its particular needs and application domain, resulting in systems that cannot shake hands with one another in an automated manner. Obviously, it is easy to embed a link in a collection database which points to the record of a similar object from another institution. But this requires knowledge of how to access the database of the other institution, what fields are used to describe the object. Once the record is found to which the resource can be linked, the URL needs to be embedded manually within the record of the original database. These actions cannot be performed reasonably for all of a library's collections items. Therefore, there is a need to think about how the linking process can be automated.[13] Libraries have amassed an enormous amount of machine-readable data about their own collections, both physical and electronic, over the past 50 years. However, this data is currently in proprietary formats understood only by the library community and is not easily reusable with other data stores or across the Web.[14]

The rise of the Web obliged libraries and other culture curating institutions to increase the pace of their standardisation efforts for metadata schemes and controlled vocabularies, which were initiated after the use of databases for cataloguing and indexing in the 1970s and 1980s. At the same time, budget cuts and fast-growing collections are currently obliging information providers to explore automated methods to provide access to resources simply because libraries are now expected to obtain and

12 K. R. Diaz, 'The Invisible Web: Navigating the Web outside Traditional Search Engines', *Reference & User Services Quarterly*, vol. 40, no. 2, 2000, pp. 131–134, http://kb.osu.edu/dspace/handle/1811/44703 [access: 23/03/2016].

13 S. van Hooland, R. Verborgh, op. cit., pp. 49–50.

14 M. Teets, M. Goldner, 'Libraries' Role in Curating and Exposing Big Data', *Future Internet*, vol. 5 (3), 2013, pp. 429–438, www.mdpi.com/1999-5903/5/3/429 [access: 23/03/2016].

provide more value out of the metadata patrimony they have been building up over decades. The current hype surrounding linked data and the Semantic Web technology underlying linked data seems to offer amazing opportunities to valorise what libraries already possess and to facilitate the creation of new metadata.[15] This relatively new generation of Web technology is designed to improve communication between people and programmes that use differing terminologies, as well as to extend the interoperability of databases, to provide tools for interacting with multimedia collections, and to provide new mechanisms for the support of "agent-based" computing in which people and machines work more interactively.[16] The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users. The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, enabling computers and people to work in a better cooperation. The first steps in weaving the Semantic Web into the structure of the existing Web have already been taken, and these developments will usher in significant new functionalities as machines become much better able to process and "understand" the data that they are merely displaying at present.[17] Whereas the original Web provides links between pages that are designed for human consumption, the Semantic Web augments this with web pages designed to contain machine-readable descriptions of Web pages and other Web resources. These documents can be linked together to provide information to the computer as to how the terms in one document relate to those in another one. To achieve this desired effect, the Semantic Web uses Web languages based on RDF (the Resource Description Framework), which go beyond the presentation capabilities of HTML and the document-tagging capabilities of XML.[18] RDF is designed to represent information in a flexible way. The generality of RDF facilitates sharing information between applications by making the information accessible to more applications across the entire Internet.[19] RDF goes beyond metadata description by providing a model for the relationships between human-generated and machine-

15 S. van Hooland, R. Verborgh, op. cit., pp. 1–2.

16 J. Hendler, op. cit., p. 520.

17 T. Berners-Lee, J. Hendler, O. Lassila, 'The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities', *Scientific American*, 01/05/2001, www.scientificamerican.com/article/the-semantic-web/ [access: 23/03/2016].

18 J. Hendler, op. cit., p. 520.

19 H. Haidarian Shahri, op. cit., p. 33.

generated (and processed) metadata and works with different types of objects or data entities. RDF can still be expressed in mark-up languages such as the aforementioned HTML and XML, but there are more available RDF serialisations. Its purpose is to enable encoding, exchange and re-use of metadata definitions and schemas. The system is flexible, as it allows each resource description community to define its own metadata elements. It also allows those communities to access existing schemas and to re-use elements that might be relevant to their data model. The namespace convention ensures that there is a unique reference back to the original definition. This system exploits the power and range of the Internet and avoids the need for a central register or repository of the data elements. As an object oriented system, RDF is based on three types of objects: resources, properties, and statements. A statement applies to a specific resource and includes a subject (the resource), the predicate (the property) and object (the value of the property). This statement syntax subject-predicate-object is known as the RDF-triple.[20]

The term "semantic search" applies when the search is performed on data stored in the RDF data model. When the data is modelled in RDF, it inherently contains explicitly typed relations or semantics and hence the use of the term "semantic search." With the semantic search, the intentions of a user can be accurately narrowed down in a more robust way than keyword search. Keyword search, popularised by the success of the Web search engines, has become one of the most widely used techniques for finding information on the Web. However, the customary indexing of keywords, as done by Web search engines, is only effective on text Web pages, also referred to as the shallow Web. It is generally accepted that the deep Web, which contains structured and semi-structured data, is significantly larger than the shallow Web. In other words, without applying the Semantic Web technology to these structured and semi-structured datasets, a considerable amount of data is still "locked away" in databases in structured and semi-structured format.[21] This state might be easily explained by the fact that, until recently, the Web has been developing most rapidly as a medium of documents for people rather than of data and information that can be processed automatically. The Semantic Web aims to make up for this. For the semantic web to function, computers must have access to structured collections of information and sets of inference rules that they can use to conduct automated reasoning. This reasoning is based on the triples of

20 D. Hayes, *Metadata for Information Management and Retrieval*, London 2004, pp. 41–42.

21 H. Haidarian Shahri, op. cit., pp. 32–34.

RDF, which form webs of information about related things.[22] With the semantic search, in the RDF model, users can iteratively refine their search, navigate through the initial results, and filter out the results (entities) that do not have the properties they are looking for. In fact, the explicit representation of properties in RDF (which does not exist in the classic database relation model supporting only the keyword search) facilitates this refinement of search results. The traditional keyword search process requires various steps, including: finding keys in tables, finding joinable attributes, generating foreign key join candidates, and removing semantically incorrect candidates. In contrast, in semantic search, the user's knowledge of the domain can be utilised effectively to navigate through sets of entities and refine the search results. Moreover, enumerating all possible candidate networks that may contribute to the results is computationally expensive, while this "user-driven" navigation in semantic search replaces the enumeration of candidate networks in keyword search.[23]

In many knowledge representation systems, there is a problem with comparing or creating relations between two (or more) knowledge databases because the former models implied that every concept or term has exclusively one place in the tree of knowledge. On the other hand, as the framework for the Semantic Web, RDF was designed for retrospective documentation of relations between initially independent terms and concepts.[24] When we want to make resources and their metadata available in a structured manner on the web, we first need to decide which of their characteristics are the most important ones to be represented. By doing so, we make an abstraction of the reality through the development of a model. In the cultural heritage context, institutions are forced to work with off-the-shelf software, since the development of a custom-built collection management system is in most cases simply not economically feasible. The drawback of working with existing software is that institutions often find themselves limited in how they can describe their objects. Vendors have a commercial incentive to develop generic software that can be sold to as many institutions as possible. This implies that collection management software already prescribes a certain explicit worldview through the use of a pre-established model. It is therefore not always possible to accommodate the specific requirements of an institution and its collections,

22 T. Berners-Lee, J. Hendler, O. Lassila, op. cit., p. 1.

23 H. Haidarian Shahri, op. cit., pp. 36–38.

24 M. Nahotko, *Metadane: Sposób na uporządkowanie Internetu* [Metadata: solution for Internet arrangement], Kraków 2004, p. 56.

leading to frustration amongst collection holders.[25] Despite the development of the Semantic Web technologies, knowledge representation is still in a state comparable to that of hypertext before the advent of the Web: it is clearly a good idea, and some very attractive demonstrations exist, but it has not yet changed the world. It contains the seeds of important applications, but to realise its full potential it must be linked into a single global system. Traditional knowledge-representation systems have typically been centralised, requiring everyone to share exactly the same definition of common concepts. But central control is stifling, and increasing the size and scope of such a system rapidly becomes unmanageable.[26] For that reason, the Semantic Web, providing information content in a form suitable for processing by machines, is achievable only by creating further interoperability layers. In order to make this possible, it is necessary to apply standards not only for the syntactic form of documents, but also for their semantic content. The RDF scheme provides the syntactics necessary for the creation of thesauri, structures, and frameworks to express the metadata describing the contents of the Web. However, for defining formal semantics for the scheme that might be used for deduction and creation of implicit relations, there is an indispensable need for building up additional elements on top of the RDF scheme. The key role in the creation of the Semantic Web, especially in the scope of knowledge representation techniques, can be assigned to semantic networks, predicate calculus, and ontologies.[27]

Ontologies is one of the new terms which in the past few years have emerged within the field of information science, but are in reality applied to older, well-known concepts whose changes in meaning have been caused simply by the entry of information technology into the world of documentation and information. Computer scientists have begun to develop computer programs without taking into account professionals in the fields of documentation and information. In some philosophical treatises, ontology is described as the study of what exists and what we assume exists in order to achieve a coherent description of reality. This description may be the key to understanding the appropriation of this term by computer scientists. There is a desire to find a parallel between "the study of what exists" – that is to say, a domain of knowledge – and "what we assume exists", or the transformation from a natural language, a reality of the chosen

25 S. van Hooland, R. Verborgh, op. cit., pp. 1–2.

26 T. Berners-Lee, J. Hendler, O. Lassila, op. cit., p. 2.

27 M. Nahotko, op. cit., p. 56.

domain, to a codified language, which is what we "assume exists", in order to "achieve a coherent description of reality".[28] Ontologies in information technology are generally defined as "representations of the distributed conceptualisation of a certain domain", where "conceptualisation" means the abstract worldview which can be expressed by metadata. Ontologies connect dictionary terms with entities identified during the conceptualisation, and make available the definitions that allow for the clarification of these terms. This way they provide an unambiguous understanding of the domain, and this understanding can then be conveyed further to human users and application systems. Ontology is a logic theory represented by the intentional meaning of formalised thesauri. One of the goals of creating ontologies is to improve the functioning of search and retrieval systems and information representation systems. Another example of the exploitation of ontologies, though still related to the first one, is their implementation in Web search-engines, which allows for extending the traditional keyword search and information retrieval by the semantic search, wherever the metadata formats allow for it.[29] Ontologies can enhance the functioning of the Web in many ways. They can be used in a simple fashion to improve the accuracy of Web searches, as the search program can look for only those pages that refer to a precise concept instead of all those using ambiguous keywords. More advanced applications will use ontologies to relate the information on a page to associated knowledge structures and inference rules.[30] The Web Ontology Language (OWL) is a semantic markup language, meant for knowledge representation. OWL is used for authoring, publishing, and sharing ontologies on the World Wide Web. The OWL language has a formal semantics, derived from Description Logics, and uses RDF/XML serialisation. It is endorsed by the World Wide Web Consortium (W3C) for the Semantic Web and is developed as a vocabulary extension of RDF. OWL has many versions and provides simply a framework for expressing various vocabularies, which can constitute a variety of differing ontologies.[31]

New software tools have been developed for mapping and linking terms between different ontologies, for using ontologies in the markup of Web

28 E. Currás, *Ontologies, Taxonomies and Thesauri in Systems Science and Systematics*, Oxford 2010, pp. 17–19.

29 M. Nahotko, op. cit., pp. 56–57.

30 T. Berners-Lee, J. Hendler, O. Lassila, op. cit., p. 3.

31 M. S. Mir, *Semantic modeling of requirements: leveraging ontologies in systems engineering*, dissertation, University of Arkansas at Little Rock, 2012, pp. 59–60, http://dl.acm.org/citation.cfm?id=2518879 [access: 23/03/2016].

sites, research publications and databases, and for capturing semantic metadata about images and other multimedia objects. Furthermore, new search technologies have been under development to exploit ontological and other Semantic Web technologies as well as to extend the capabilities of Semantic Web languages in order to allow for more complex information to be expressed (for example, representing how a particular process might change over time, or how a set of Web-accessible programs could be automatically combined). Of particular note are some of the first demonstrations of Semantic Web "agents" that can integrate the information from Web pages and databases, and then pass it to programs for analysis and query processing.[32] By "agents" we understand the many programs that collect Web content from diverse sources, process the information and exchange the results with other programs, thus bringing the real power of the Semantic Web to its full potential. The effectiveness of such software agents increases exponentially as more machine-readable Web content and automated services (including other agents) become available. The Semantic Web promotes this synergy: even agents that were not expressly designed to work together can transfer data to one another when the data come with semantics. Many automated Web-based services already exist without semantics, but other programs such as agents have no way to locate one that will perform a specific function. In contrast, the Semantic Web is more flexible. The consumer and producer agents can reach a shared understanding by exchanging ontologies, which provide the vocabulary needed for discussion. The agents can even develop new reasoning capabilities when they discover new ontologies. Semantics also make it easier to take advantage of a service that only partially matches a request.[33] The implementation of the RDF model in the open and distributed context of the web is based upon its capability to issue identifiers for subjects, predicates and objects, which can be freely reused. Software is then able to interpret this information, because the identifiers create unique meaning, as opposed to the names of columns in databases or elements in XML, which only have local significance and change from application to application.[34] Nevertheless, it should be stressed that the Semantic Web is not just about putting data on the web in the appropriate format. It is about making links, so that a person or machine can explore the web of data. With linked data, other, related

32 J. Hendler, op. cit., p. 521.

33 T. Berners-Lee, J. Hendler, O. Lassila, op. cit., p. 3.

34 S. van Hooland, R. Verborgh, op. cit., p. 45.

resources can be easily found. Linked data is essential to actually connect the Semantic Web.[35] In order to move forward with a machine-readable web, Berners-Lee, one of the founders of the World Wide Web, came up with the linked data principles. These four rules specify a simple way to format data so that it can be interpreted by software. The rules can be summarised in the following way (and are numbered accordingly):

- necessarily using URIs (Uniform Resources Identifiers) as names for things,
- preferably using HTTP URIs, in other words – URLs (Uniform Resource Locators) so that they can be looked up via the standard http protocol,
- providing useful information for the URIs that have been looked up, using the Semantic Web standards (RDF, SPARQL),
- including links to other URIs so that more resources can be discovered.

The process of looking up more information about a subject through its URL is called dereferencing.[36] Although there is no strict rule about it, dereferencing with authoritative data sources such as libraries is considered a good practice, since authoritative data are considered the most accurate and have been vetted according to official rules and policy. The data have a known accuracy and lineage, and can be verified and certified by data stewards in the authoritative source.[37] There is already strong evidence that, once exposed, library data is useful to other communities and is accessed and repurposed. Because librarians have invested so much time and effort into authoritatively describing resources, their creators, and subjects covered, this data serves a valuable role in connecting the many aspects of people, items, places, events, organisations and concepts into a meaningful knowledge graph.[38] Aside from the explicit encoding of semantics, the advantage of the Semantic Web and the RDF model, upon which it is built, is also the global referencing of entities on the entire web, which does not exist in the classic relational model. The global referencing of entities in RDF is vital to facilitating the interoperability and, consequently, aggregation and re-use of knowledge across organisational

35 T. Berners-Lee, 'Linked Data: Design Issues', 27/07/2016, www.w3.org/DesignIssues/LinkedData.html [access: 23/03/2016].

36 S. van Hooland, R. Verborgh, op. cit., pp. 45–47.

37 D. Stage, 'Authority and Authoritative Data: A Clarification of Terms and Concepts', *Fair & Equitable*, February 2009, pp. 13–16, www.iaao.org/uploads/Stage.pdf [access: 23/03/2016].

38 M. Teets, M. Goldner, op. cit., p. 437.

boundaries. It must be noted, however, that facilitating interoperability across distributed and heterogeneous databases is quite difficult, partly due to the lack of such a referencing mechanism.[39]

It can be concluded that linked data technologies, a term used often interchangeably with Semantic Web technologies, constitute a primary enabler for interoperability at scale on the Web. The use of linked data in a larger knowledge graph, and its application in Big Data, is not just theory. It provides immediate and tangible benefits to the consumers of Big Data being managed.[40] However, linked data principles are often misunderstood and need to be implemented in a thoroughly considered manner. Linking data presents tremendous challenges with regard to the quality of library metadata, and so it is fundamental to develop a critical view and differentiate between what is feasible and what is not.[41]

## Metadata and Interoperability

Consequently, research efforts in the Semantic Web and Linked Data have significantly fuelled interest within the database community regarding the problems raised by using data sources on the web. The rapid growth of the Internet and the Web has necessitated the creation of more principled mechanisms to facilitate semantic interoperability and the querying of data across organisational boundaries. Despite many years of work on semantic interoperability, this problem is still open. Moreover, it has acquired a new urgency, now that physical and syntactic interoperability barriers have for the most part been removed. Physical interoperability between systems has been solved with the advent of hardware standards such as Ethernet and with protocols such as TCP/IP and HTTP. The syntactic interoperability between systems has been largely solved by agreeing on the syntactic form of the data that we exchange, particularly with the advent of XML (and more recently the RDF standard for Semantic Web), which has been briefly introduced above. For semantic interoperability between systems, we not only need to know the syntactic form (structure) of the data, but also the intended meaning of the data.[42] It might be said that there are two contexts for metadata and interoperability: metadata as a tool to enable the exchange of information between

39 H. Haidarian Shahri, op. cit., p. 39.

40 M. Teets, M. Goldner, op. cit., pp. 431–433.

41 S. van Hooland, R. Verborgh, op. cit., pp. 2–3.

42 H. Haidarian Shahri, op. cit., pp. 12–13.

interoperating systems, and the interoperability of metadata schemas themselves, which can facilitate systems' interoperability.[43]

It may already be apparent, but it is still worth stressing that the term "interoperability" is closely related to metadata issues. Interoperability itself is a capacity of two or more systems (or departments of these systems) to exchange information and to be able to process this information without additional operations (either manual or automatic) within any of these systems (or their departments).[44] Interoperability depends on the exchange of metadata between systems to establish the nature of the data being transferred and the way in which it should be handled. Metadata acts as an enabler of information and data transfer between systems, and as such it is a key component in interoperability. In order to allow for software applications that have been designed independently to pass data between them, a common framework for describing the data being transferred is needed so that each application "knows" how to handle the data in the most appropriate manner. This might be at the level of distinguishing between different data formats or between different vocabularies.[45]

The evolution from an unstructured narrative to a highly structured representation of metadata requires the development of schemas in order to make the metadata interoperable. By slicing up unstructured descriptive narratives into well-structured fields, we need to render the meaning of the different fields (also called attributes) explicit by documenting them in a schema. By structuring and atomising metadata fields we make them more machine-interoperable, but at the same time we become more and more reliant on the schemas when we need to interpret either our own metadata or those of someone else. It is precisely in this context that linked data need to be understood. Through the adoption of a radically simple data model, abstractions can be made of the traditional XML and database schemas we had to use in the past to interpret and re-use data.[46] Much of the research on metadata is related to the creation of standards and their adoption. It is necessary to integrate the new standards more tightly with those already existing, and with their applications. The implementation of these requirements leads to the development and acquisition of models and meta-models. A metadata model should be designed to be general and abstract, otherwise its usefulness will be limited. On the

43 D. Hayes, op. cit., p. 159.

44 M. Nahotko, op. cit., p. 48.

45 D. Hayes, op. cit., pp. 12–16.

46 S. van Hooland, R. Verborgh, op. cit., pp. 12–13.

other hand, it is necessary to ensure easy representation of the model by means of syntactic structures (such as XML or RDF), and relational or object database models. It should have its own coherent and unified dictionary, and at the same time it should be easily adapted to other existing systems – semantic and syntactic – used in existing metadata schemes.[47]

The advantages of RDF, the data model underlying the linked data vision, can only be fully understood in the context of previous data models. Up to the beginning of this century, sharing data between different databases was a very tedious process. Meta-markup languages, and XML in particular, have been used since 2000 in order to facilitate the exchange of structured data on the web. XML offered a standardised syntax for the automated exchange of structured data, but the actual use and interpretation of such data can still be troublesome. The meaning of the elements and attributes of the XML files need to be defined in a schema. The interpretation of the schema remains a barrier for the automated consumption of data across information systems on the Web. It is exactly here that RDF comes in. By adopting a data model which embodies the meaning of the data in its most essential and stripped down form, the need for an outside schema to interpret and re-use the data no longer exists.[48] The problem of metadata exchange faced by special and digital collections curators using the Dublin Core scheme might serve here as an example. Native metadata records of such collections are usually rich in meaning in their own environment, but lose their robustness in the aggregated environment due to mapping errors and to misunderstanding and misuse of the Dublin Core scheme elements. This phenomenon is closely related to a sharable metadata issue, namely that metadata may be of high quality within its local context, but may be compromised when taken out of this context for various reasons. Without context, useful local information may be lost, become insignificant, or become ambiguous and cause confusion to users in aggregator environments. Since no single metadata standard works for every digital collection, it is inevitable for collection curators to develop and use locally defined unique fields for collections in their local environments. The challenge then is to support metadata aggregation and other forms of interoperability by maintaining context to the maximum possible degree, even while normalising metadata records for sharing with others.[49]

47 M. Nahotko, op. cit., pp. 86–88.

48 S. van Hooland, R. Verborgh, op. cit., pp. 14–15.

49 M.-J. Han, C. Cho, T. W. Cole, A. S. Jackson, 'Metadata for Special Collections in CONTENTdm: How to Improve Interoperability of Unique Fields Through OAI-PMH', *Journal of Library Metadata*, vol. 9, 2009, pp. 213–238, www.ideals.illinois.edu/bitstream/handle/2142/15415/WJLM.pdf?sequence=2 [access: 23/03/2016].

Importing metadata raises issues about the choice of schema. A number of registries and crosswalks are available which can help with the selection of appropriate schemas and also, therefore, of potential sources of metadata, yet as convergence takes place between different domains of activity, new challenges arise for establishing interoperability between domain-specific metadata schemas and application profiles. Previous research has suggested that a new Web metadata architecture based on the best features of RDF and XML can enhance the interoperability between application profiles.[50] The reason for this is that by adopting an extremely simple data model consisting of triples, data represented in RDF become schema-neutral. An RDF triple consists of a subject, a predicate, and an object. This allows for maximum flexibility. Any resource in the world (the subject) can have a specific relationship (the predicate) to any resource in the world (the object). There is no limit to what can be connected with what. By simplifying the data model to a maximum, the whole semantics is made explicit by the triple itself. By doing so, there is no longer a need for a schema to interpret the data. Within the world of databases and XML, only the data conforming to the rules defined in the schema may exist and be encoded in the database or XML file. With RDF, you only make statements about facts you know, but these statements might interact with statements made outside your information system. This data model allows for heterogeneous data to connect and interact. However, the fact that the model is schema-neutral does not mean that no schema-related issues remain. Any piece of data still needs to be expressed in a certain vocabulary, and each vocabulary has its own way of expressing things.[51] An important feature of the Web architecture is defining the main concepts related to linked resources and metadata. The meaning of these relations is worth consideration. Sometimes these relations are specific to the Web, as defined in its architecture, or are important for the protocols used. In other cases, the importance of relations and attributes is part of another specification, project, or application, and must be defined elsewhere. Therefore, a set of such relations and attributes should be easily expanded, and in a decentralised way. Thus determined requirements, defined in this way, are met by specifying the location using the URL that is appropriate for the definitions of the attribute names.[52]

50 D. Hayes, op. cit., p. 162.

51 S. van Hooland, R. Verborgh, op. cit., p. 44.

52 M. Nahotko, op. cit., pp. 35–36.

The cultural heritage sector, including museums, libraries, and archives, has a long history of using metadata to describe intellectual and artistic works, and is now using its long traditions of cataloguing to create metadata for digital resources as well.[53] Libraries have been working for decades with the MARC format, which is an electronic file format created in the 1960s to represent flat files containing bibliographic data.[54] MARC (which stands for Machine Readable Cataloguing) had been developed to provide a method to encode bibliographic and authority data so it could be moved between systems. However, it has only been used in the library community and is not appropriate for use in today's Semantic Web.[55] Since the MARC standard has evolved into a family of national standards (MARC 21 used by the National Library of Poland is one of these standards that emerged from the one known formerly as USMARC), its value as a standard had already decreased before the rise of Semantic Web technology. The MARC standard forms the basis for shared library cataloguing and for the exchange of data between different library management systems. The international standard ISO 2709:1996 defines the record syntax for MARC with tags and content defined by national cataloguing authorities.[56] Organising the authoritative descriptions of library objects is a well-understood concept in library cataloguing. However, as already mentioned, libraries have primarily interoperated by exchanging text strings encoded in industry-proprietary MARC record formats. Library systems have evolved collectively to consume, manage and reproduce these text strings. Authoritative text strings have been stored and managed separately, and are somewhat loosely connected through applications to item descriptions. The transition to effectively operating with data at scale requires much more focus on accessible structures, persistent identifiers, and comprehensive modelling of the data under management. Linked data is the formalised method of publishing this structured data so that large data repositories can be interlinked using standard Web protocols. In order to manage the authority of text strings, the organisation of top-level entities relevant to the specific needs of libraries must be taken into account. This organisation is often referred to as an "upper ontology" or, more recently, as entities in a knowledge graph.[57]

53 D. Hayes, op. cit., p. 56.

54 S. van Hooland, R. Verborgh, op. cit., p. 4.

55 M. Teets, M. Goldner, op. cit., p. 431.

56 D. Hayes, op. cit., p. 57.

57 M. Teets, M. Goldner, op. cit., p. 432.

While for almost three decades there have been forecasts about the imminent end of its use, the MARC standard is still being implemented in ever-newer generations of information systems. This is perhaps inevitable, as millions of bibliographic records have been created in this format, yet it undergoes various modifications, some of them aiming at reconciling this standard with the needs presented by the Web of data.[58] While there has recently been good progress in understanding the character and importance of shareable metadata, comparatively less effort has been expended to date in examining the degree to which collection curators still locally develop and customise metadata standards for use in their local environments, and how these customised metadata standards are mapped to standard metadata schemes for dissemination to metadata aggregators. It must be stated that despite some shortcomings that collections curators might encounter, it is a good practice to comply with the chosen metadata standard, because locally defined, unique fields – which potentially have substantial contextual information – could impede the interoperability of metadata, since the contextual information in such fields as implemented could not be or was not being mapped in ways which would facilitate interoperability.[59]

It is another acknowledged truth that it would be a huge waste not to re-use and value the work performed by cultural heritage institutions on the development of thesauri (which in fact has been performed by them over decades preceding the current approach to metadata) in the context of linked data. The linked data movement clearly stirred a new interest in traditional controlled vocabularies. Controlled vocabularies can play a pivotal role in addressing the third and fourth linked data principles, that is to provide useful information related to the URI looked up, using the standards and inclusion of links to other URIs, so that by following these links more things can be discovered. A thesaurus expressed in the Simple Knowledge Organisation System (SKOS) is available in an RDF-based data format and can contain links to other URIs, representing for example narrower, broader, or related terms. The added value of publishing, for example, thesauri on the web is not only to create links between objects and descriptors, but also to facilitate interconnections between descriptors of multiple thesauri, and to automatically map uncontrolled keywords as much as possible to a thesaurus published in the SKOS format.[60]

58 M. Nahotko, op. cit., p. 134.

59 M.-J. Han, C. Cho, T. W. Cole, A. S. Jackson, op. cit., pp. 215–216.

60 S. van Hooland, R. Verborgh, op. cit., pp. 111–112.

The emergence of linked data for libraries began in fact with the Library of Congress' publication of LCSH (Library of Congress Subject Headings) in SKOS. As more and more RDF-based metadata become available, a lack of established best practices for vocabulary development and management in the Semantic Web world is leading to a certain level of vocabulary chaos. Metadata element sets and value vocabularies, along with datasets, are contexts recently defined and scoped for the linked data of archives, libraries, and museums. Until recently, vocabularies were considered to be tied tightly to particular domains and applications. In the library world, most vocabulary development has been taking place in the context of MARC 21, and similar development trajectories have occurred within other domains of practice.[61] The difficulties of implementing the full-blown Semantic Web, based on the approach towards formalisation of the natural language, led to a renewed appraisal of traditional controlled vocabularies, for which the capabilities to improve precision and recall within an information retrieval context remain relevant. From a purely pragmatic point of view, it was realised that existing vocabularies should be re-used, instead of developing new ones. The growing need to re-use and exchange controlled vocabularies in the context of the Web has given rise to standardisation efforts, of which SKOS has been so far the most successful.[62]

Reconciling metadata created in different environments is a major challenge, and significant effort has been devoted to mapping equivalent metadata elements between different metadata schemas. These mappings can be displayed, and are known as crosswalks. They can be used within systems to perform transformations between metadata objects. In the area of bibliographic standards, FRBR (Functional Requirements for Bibliographic Records) provides a model for bibliographic data that can foster the creation of crosswalks between schemes. The Library of Congress has initiated some work on this to provide a mapping of the MARC data elements onto the FRBR model and AACR2 cataloguing rules, which has since resulted in adopting new cataloguing rules in the form of the new standard for descriptive cataloguing RDA (Resource Description and Access), to which we will return in further parts of this paper. Crosswalks have been published between many major metadata schemas. They are a starting

61 G. Dunsire, C. Harper, D. Hillmann, J. Phipps, 'Linked Data Vocabulary Management: Infrastructure Support, Data Integration, and Interoperability', *Information Standards Quarterly*, vol. 24 (2/3), 2012, pp. 5–13, www.niso.org/apps/group_public/download.php/9411/FE_Dunsire-etal_VocabMgmt_isqv24no2-3.pdf [access: 23/03/2016].

62 S. van Hooland, R. Verborgh, op. cit., p. 128.

point for assessing the suitability of data sources for import. Published crosswalks provide a means of analysing steps for necessary data conversion. The proliferation of metadata standards developed by different communities of interest means that there is a significant danger of not being able to exchange metadata. One response to the growth in the number of metadata standards has been the development of metadata registries. The European Union and the International Federation of Library Associations and Institutions (IFLA) have developed information sources and registries for the exchange of information concerning the range of metadata standards and activities that are currently available. [63] The OMR (Open Metadata Registry), which is one of the most active ones, was built as a free, open service, and among its most important functions is the ability to provide detailed versioning of changes at every level. It has been used extensively in the library community, now hosting the vocabularies of RDA (Resource Description and Access), ISBD (International Standard Bibliographic Description) and the FR family of models (Functional Requirements for Bibliographic Records/Authority Data/Subject Authority Data) developed by IFLA, and the experimental version of MARC 21 in RDF (such as MARC 21rdf.info vocabulary). The OMR is now engaged in a significant redevelopment effort, focused on vocabulary mapping. Both the comprehensive efforts to model the rich depth of MARC 21, RDA, and ISBD, and the more selective exposure of key information from the data encoded by means of these standards using more common web vocabularies, are important aspects of current experimentation in linked bibliographic data.[64]

The technique of vocabulary mapping, or alignment, attempts to create connections between existing controlled vocabularies in order to establish links between objects belonging to different collections, which have been indexed and catalogued with the help of different vocabularies. Two heterogeneity problems stand in the way of vocabulary interoperability: representational heterogeneity stemming from the fact that different types of vocabularies contain different structural elements, and conceptual heterogeneity, which lies in different vocabularies referring to the same concept but with different names or labels, or differences regarding the hierarchical level. The problem of representational heterogeneity might be solved by having the vocabularies in a structured, unified format such as SKOS, but the problem of conceptual heterogeneity between vocabularies still needs to be dealt with. In another approach to reconciling metadata records, the

63 D. Hayes, op. cit., pp. 158–161.

64 G. Dunsire, C. Harper, D. Hillmann, J. Phipps, op. cit.

terms used within metadata records, typically found in a metadata field such as "Keywords", can be reconciled with existing and well-established vocabularies. In this approach, instead of performing a lengthy and complex mapping process between vocabularies, the reconciliation process bypasses that operation by checking whether a used keyword appears in the external vocabulary considered as the reconciliation source. Different methods, such as a stemming algorithm, allow for augmentation of the string matching process.[65] It must be noticed here, however, that this latter approach might return erroneous results caused by the lack of proper contextualisation for the data due to reasons of ambiguity. This problem will be discussed further in regard to the National Library Descriptors project, together with ways in which the creators of the project have approached this issue. It must be taken into account that contextualisation is one of the five basic qualities of the hyper-informational knowledge management model, where it is understood as the principle of determining relationships between concepts selected in a knowledge domain, which are mapped in the system directly in the structure of the representation of these fragments, containing both expressions identifying these concepts and expressions identifying their context.[66] It is also worth noticing at this point that both approaches to the reconciliation of vocabularies are facilitated by applying a certain level of simplicity to the vocabulary created, and by the possible usage of natural language terms, both of which were also prerequisites for the adoption of the National Library Descriptors project.

Controlled vocabularies, when used for indexing a collection of documents, allow for greater precision and recall during the search and retrieval process within an information system. In the context of linked data, controlled vocabularies also allow for connections to be created between collections. The difficulties encountered in attempts to create full-blown ontologies have created a new opportunity for controlled vocabularies in the linked data environment. Traditional vocabularies have an important role to play for the realisation of the third and fourth linked data principles by offering URIs that provide useful information accessible in a standardised format and that contain links to other URIs. Many vocabularies have already issued unique identifiers for their terms under the form of URLs. The use of SKOS has paved the way to share and distribute thesauri more easily, since SKOS, as a standard format for the representation of relations between terms, has helped over the past few years to solve the issue of

65 S. van Hooland, R. Verborgh, op. cit., pp. 132–134.

66 M. Nahotko, op. cit., p. 37.

representational heterogeneity to which the third linked data principle refers. The fourth linked data principle refers to the provision of extra URIs, which would help to discover extra information, introducing in turn the more challenging issue of providing links between vocabularies.[67] This problem has been one of the key issues underlying the adoption of the National Library Descriptors project, which aims at enhancing the National Library metadata by providing additional relations and atomising vocabulary elements so that they can be easily mapped.

## National Library Descriptors

The development of element sets and value vocabularies for RDA by the Library of Congress has taken place in an open environment, with benefits for both maintainers and consumers. The progress of such work and feedback on it are easily monitored by colleagues and other interested parties. The development of the RDA namespace has immediately stimulated the IFLA communities to consider the potential use of their own standards in the Semantic Web, as RDA is based on the FR (Functional Requirements) models family.[68] Since the National Library of Poland was utilising different cataloguing rules, namely the above mentioned ISBD standard (International Standard Bibliographic Description), adopting the full RDA methodology seemed cost-inefficient and unnecessary, especially in areas not covered by ISBD, where using the RDA namespace in the ISBD standard records has been allowed. The process of adding RDA cataloguing elements to a non-RDA bibliographic record (e.g. a bibliographic record not coded as catalogued according to RDA), either manually or via machine manipulation, is called hybridisation. The result of hybridisation is a hybrid record, rather than an RDA record. This hybridisation might be achieved by editing records created under non-RDA standards. It may be accomplished by following guidelines for specific fields, and, in some cases, by adding new MARC 21 fields (such as 336, 337, 338 or relationship designators), as long as the bibliographic integrity and identity is not affected. This allows cataloguers to make use of existing non-RDA records and minimises instances where additional help is needed. Adding RDA elements to non-RDA records allows libraries to reap the benefits of clarity, additional fields, and access points, but does not impact the bibliographic identity of the record.[69]

67 S. van Hooland, R. Verborgh, op. cit., pp. 135–136.

68 G. Dunsire, C. Harper, D. Hillmann, J. Phipps, op. cit.

69 Library of Congress, 'Report of the PCC Post-Implementation Hybrid Bibliographic Records Guidelines Task Group', 15/10/2012, www.loc.gov [access: 28/03/2016].

As this approach was adopted at the National Library of Poland, the decision was made to start using some of the MARC 21 standard fields that were not necessarily introduced to the MARC format along with the adoption of the RDA cataloguing standard but were partly inspired by this shift in approach to cataloguing and added to the MARC standard during the years preceding the decision to adopt the RDA cataloguing standard by the Library of Congress. Adding these fields to the set of already used MARC tags allows for much more precise definition of the attributes assigned to specific headings (i.e. entities, as defined in the National Library Descriptors vocabulary). It aims at creating additional relations between entities and attributes expressed by the National Library of Poland's controlled vocabulary, which will be presented later. For the Authority Records, the following set of MARC 21 standard fields has been added, previously unused at the National Library of Poland:[70]

| MARC 21 Field Tag: | When added to the standard: |
|---|---|
| 024 Other Standard Identifier (Repeatable) | New: 2003 |
| 034 Coded Cartographic Mathematical Data (Repeatable) | New: 2006 |
| 043 Geographic Area Code (Non-Repeatable) | |
| 045 Time Period of Content (Non-Repeatable) | |
| 046 Special Coded Dates (Repeatable) | New: 2009 |
| 368 Other Attributes of Person or Corporate Body (Repeatable) | New: 2011, Redefined: 2012 |
| 370 Associated Place (Repeatable) | New: 2009 |
| 371 Address (Repeatable) | New: 2009 |
| 372 Field of Activity (Repeatable) | New: 2009 |
| 373 Associated Group (Repeatable) | New: 2009, Redefined: 2011 |
| 374 Occupation (Repeatable) | New: 2009 |
| 375 Gender (Repeatable) | New: 2009 |
| 376 Family Information (Repeatable) | New: 2009 |
| 377 Associated Language (Repeatable) | New: 2009 |
| 378 Fuller Form of Personal Name (Non-Repeatable) | New: 2011 |
| 380 Form of work (Repeatable) | New: 2010 |
| 385 Audience Characteristics (Repeatable) | New: 2013 |
| 386 Creator/Contributor Characteristics (Repeatable) | New: 2013 |
| 388 Time Period of Creation (Repeatable) | New: 2014 |

70 Library of Congress, 'MARC 21 Format for Authority Data', www.loc.gov/marc/authority/ [access: 10/04/2016].

For the Bibliographic Record, the following MARC 21 fields have been added to the existing set:[71]

| MARC 21 Field Tag: | When added to the standard: |
|---|---|
| 336 Content Type (Repeatable) | New: 2009 |
| 337 Media Type (Repeatable) | New: 2009 |
| 338 Carrier Type (Repeatable) | New: 2009 |
| 380 Form of Work (Repeatable) | New: 2010 |
| 381 Other Distinguishing Characteristics of Work or Expression (Repeatable) | New: 2010 |
| 385 Audience Characteristics (Repeatable) | New: 2013 |
| 386 Creator/Contributor Characteristics (Repeatable) | New: 2013 |
| 388 Time Period of Creation (Repeatable) | New: 2014 |
| 658 Index Term-Curriculum Objective (Repeatable) | |

The last one on this list, the 658 MARC tag field, was not in fact a recent addition to the MARC format. However, the 658 MARC tag field has been newly adopted by the National Library of Poland within the National Library Descriptors project in order to index the area or field of knowledge after confirming it with the Library of Congress for the 'content objective'. Since this has the special meaning of providing additional context to other indexed information in the bibliographic record, it is worth discussing further what will be indicated more elaborately by presenting some examples of how the information indexed in the 658 MARC tag field may add additional meaning necessary for disambiguation and proper entity recognition.

The desired objectives of the National Library Descriptors project could not be achieved just by appending additional MARC 21 fields to the previously used set of tags. Another significant reform is also a prerequisite to the project, and this concerns the transformation of the controlled vocabulary used at the National Library of Poland, which is called the 'National Library Descriptors' in its desired, refined form, and from which the whole project's name is in fact derived. It has been previously explained how more atomised and simplified controlled vocabularies might facilitate building the relational structure of the Web by increasing the semantic interoperability. Within the described project, the National Library Descriptors are supposed to provide a controlled vocabulary for

71 Library of Congress, 'MARC 21 Format for Bibliographic Data', www.loc.gov/marc/bibliographic/ [access: 10/04/2016].

name headings, corporate headings, meetings, series, and work uniform titles, as a single authority file shared by the entities which were earlier divided into name and subject authority files. This major step towards simplification aims also at clarification of the relational structure in that it allows us to avoid the unnecessary duplicate relations and to focus on building additional relations, where they actually provide more semantic value or – speaking in simpler terms – more meaning. Another important and radical change within the controlled vocabulary structure is the fact that for both authority and bibliographic records (in the 6XX fields) the National Library will stop using subdivisions, or more precisely the subfields x, z, y, as subfield "v" has already fallen into disuse in past years. All values that were earlier recorded as general subdivisions "x", geographic subdivisions "z", and chronological subdivisions "y" will be recorded as subjects in the 650 fields, content objectives in the 658 fields or simply as geographical or chronological terms (fields 651 and 648 respectively), depending on the actual semantics of the transformed subdivision. The transformation of subdivisions constitutes the major challenge for the project, especially as the earlier version of the controlled vocabulary used them in a very elaborate way, which excludes a simple approach to mapping its values to other vocabularies. Consequently, no authority records for subdivisions will be used (eradicating the whole 18X field group). Since the same controlled vocabulary is supposed to denote the attributes recorded in the added indexed MARC tag fields, it also automatically creates a foundation for a more elaborate relational structure in the native relational database of the National Library of Poland. Unlike the Library of Congress, which uses various controlled vocabularies for the descriptions stored in its dataset, the National Library of Poland is constrained to its native controlled vocabulary due to the specifics of the vocabulary's prerequisites, as well as the specifics of the Polish language. This situation has many advantages, such as avoiding duplicate entities and ambiguous attributes in the native dataset.[72]

The actual benefits from all these challenging endeavours lie in the fact that, once the transformation of the current dataset is complete, it will consist of endorsed entities with atomic attributes expressed in a simple controlled vocabulary, giving broad scope for possibilities to interconnect it with other datasets available on the Web of data. For instance, an attri-

72 Z. Żurawińska, 'Deskryptory Biblioteki Narodowej w systemie bibliotecznym: rekord wzorcowy i bibliograficzny' [National Library Descriptors: authority record and bibliographic record], 21/04/2015, www.bn.org.pl/download/document/1429787847.pdf [access: 14/04/2016].

bute recorded in the 374 MARC 21 field with the value "Aktor" (which is the Polish language equivalent for the word "actor") might be automatically mapped to the same semantic value of the class in the widely used schema.org vocabulary, expressed in the simple N-Triples RDF serialisation:

_:Aktor <http://www.w3.org/2002/07/owl#sameAs> <http://schema.org/actor>.

where "_:" stands for the as yet undefined part of a URL pointing at the domain and/or directory where the National Library Descriptors dataset would be located. The "http://www.w3.org/2002/07/owl#sameAs" is a URI which stands for and predicates expression of the "same as" statement, and "http://schema.org/actor" is a URI pointing at the "actor" property in the schema.org vocabulary.

The same example in Turtle RDF serialisation could be presented as follows:

@prefix dbn: <http://alpha.bn.org.pl/dbn> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix schema: <http://schema.org> .
dbn:Aktor owl:SameAs schema:actor .[73]

where "dbn" stands for the hypothetical alias of the hypothetical URL "http://alpha.bn.org.pl/dbn" pointing at the domain and/or directory where National Library Descriptors dataset could be located.

By means of such a relation, the Person entities with the attribute "Aktor" would no longer be linked solely to the "Actor" class in the native dataset, but also to the equivalent property in the external standard vocabulary of

73 Choosing the "dbn" shortcut as the hypothetical alias of the hypothetical URL "http://alpha.bn.org.pl/dbn" is not accidental, as "dbn" stands for "Deskryptory Biblioteki Narodowej", which is the Polish equivalent to "National Library Descriptors". However, as mentioned above, neither the domain nor the directory nor the alias for the National Library Descriptors have yet been defined, thus this part of an example is purely hypothetical and presented here for explanatory purposes only. Please note, that the given example is only valid in OWL Full, and is not correct in OWL DL, what is explained further without details as differences between OWL DL and OWL Full are beyond the scope of this article. For the sake of simplicity our example is based on the assumption that individuals with the "actor" property would be equivalent to individuals belonging to the class "Actor", which is reasonable, but the whole reasoning mechanism is not presented.

schema.org or any other open vocabulary that provides properties and values corresponding to the National Library Descriptors vocabulary.

While continuing to follow the example of the schema.org vocabulary, we provide in the table below some examples of these vocabulary properties, allowing for mapping of the newly added MARC 21 fields in the Authority Records. In order to understand the real benefits of the project, it is crucial to note that the content of the table below represents only some selected examples of links to only one standard external vocabulary, and that there are many more such possibilities, since there exist more available predicates and vocabularies.[74] However, none of these links would have been possible if the effort of the National Library Descriptors project had not been undertaken.

| **MARC 21 Tag:** | **predicate** | schema.org **vocabulary property** |
|---|---|---|
| 024 | owl:equivalentProperty | schema:sameAs |
| 034 | owl:equivalentProperty | schema:geo |
| 046 | owl:equivalentProperty | schema:birthDate, schema:deathDate; schema:foundingDate |
| 368 | owl:equivalentProperty | schema:AdministrativeArea |
| 370 | owl:equivalentProperty | schema:birthPlace, schema:deathPlace, schema:foundingPlace, schema:location |
| 371 | owl:equivalentProperty | schema:address |
| 372 | owl:equivalentProperty | schema:industry |
| 373 | owl:equivalentProperty | schema:affiliation |
| 374 | owl:equivalentProperty | schema:jobTitle |
| 375 | owl:equivalentProperty | schema:gender |
| 376 | owl:equivalentProperty | schema:relatedTo |
| 377 | owl:equivalentProperty | schema:inLanguage |
| 385 | owl:equivalentProperty | schema:audienceType |
| 386 | owl:equivalentProperty | schema:nationality |
| 388 | owl:equivalentProperty | schema:dateCreated |

74 In fact, it should be stressed that in order to achieve the most complete mapping of the National Library Descriptors dataset to the existing vocabularies, many vocabularies must be used. The simple reason is that most of these vocabularies were created with different purposes than to express bibliographic metadata and most of them will therefore not provide equivalent elements for the whole National Library Descriptors metadata elements set but only parts of it, making combining more vocabularies necessary.

We use here the property owl:equivalentProperty to tie together properties as parts of another ontology, that is to indicate that a particular property in one ontology is equivalent to a class or property in a second ontology. In a similar fashion, the property owl:equivalentClass is used to indicate that two classes have precisely the same instances, where classes simply denote sets of individuals and are not individuals themselves (as in the OWL DL). The "owl:sameAs" statement, used in the example above, is similar to the owl:equivalentClass property we use for classes, but declares two individuals to be identical. A typical use of sameAs would be to equate individuals defined in different documents, as part of unifying two ontologies. However, in OWL Full "sameAs" may be used to equate anything: a class and an individual, a property and a class, etc., and causes both arguments to be interpreted as individuals.[75] Explaining all the possibilities of how the National Library Descriptors metadata might be linked to other vocabularies and to other datasets is far beyond the scope of this paper. We simply want to draw attention to the fact that there are many vocabularies to become linked with, and that there are more available ontologies to choose from in order to have the National Library Descriptors datasets mapped to another. For instance, an attribute recorded in the 374 MARC 21 field, as in the above example for the "Aktor" value, might be mapped to the same value of the class not only in the schema.org vocabulary, but for example to the class dbo:actor in the DBpedia ontology. In this case however, it would be rather the mapping of the class "Actor" recorded in the 150 MARC 21 field subject heading, as we should preferably use the owl:equivalentClass property for classes. Yet the relation between the person entity and the class "Actor" which we record as an attribute is easy to maintain by using, for example, the "rdf:type" predicate. In respect to all the possibilities it is also worth noting that we are not just restricted to mappings of the "equivalent" or "sameAs" properties as predicates, since the SKOS vocabulary, to name just one, allows us to link the entities by indicating broader, narrower or related terms with the properties skos:broader, skos:narrower, skos:related, and there are many more options even within the SKOS data model, not to mention other ones.[76]

We have already mentioned that the 658 MARC tag field – which instead of simply tagging the subject of the catalogued material, provides

75 W3C, 'OWL Web Ontology Language Guide: W3C Recommendation', 10/02/2004, www.w3.org/TR/owl-guide/ [access: 26/03/2016].

76 W3C, 'SKOS Simple Knowledge Organization System Namespace Document – HTML Variant', 18/08/2009, www.w3.org/2009/08/skos-reference/skos.html [access: 26.03.2016].

the context for the recorded subjects headings – has a special meaning for enriching a record's metadata. In order to illustrate exactly how we can benefit from this additional metadata information, we need to pay attention to yet another process to be observed in libraries nowadays, specifically: the creation of digital libraries and the problem we currently face of trying to grasp and make sense of the tremendous numbers of digital objects being created. Confronted with the flood of digitised resources, the Library Information Science (LIS) community and cultural institutions have in parallel undergone an important shift in thinking. Originally developed for the press and media industry, named-entity recognition (NER) is now increasingly considered as a logical extra step to be applied after the conversion of a scanned textual document from an image format into an indexable text document. The practices which are currently emerging from both the LIS and digital humanities communities are complementary and help to understand the extent to which NER can offer added value to metadata management.[77] Entity retrieval performed by the NER process is a search task that finds the entity objects in unstructured noisy documents like HTML pages or text documents according to users' information needs. Since the second half of the 1990s, many approaches to entity extraction have been examined and they can be categorised into four classes: rule-based methods, and supervised, semi-supervised or unsupervised learning methods. The tools of NER, or entity identification or extraction, locate and classify atomic elements in texts within predefined categories, such as the names of persons, locations, and times.[78] The NER technique aims at identifying sequences of words in a text that correspond to a predefined taxonomy of entities, such as people, organisations, and locations. As with the related technology of part-of-speech tagging, most approaches to NER attempt to label each word in a sentence with its appropriate class. For part-of-speech tagging, these classes are syntactic classes, such as adjectives, prepositions, common nouns, etc. In NER, the taxonomy of entities is usually small, and nonentities are often given a separate "not-an-entity" tag.[79] It is however possible to extract the entities not only by applying the NER approach to the unstructured text documents, but also to the structured data embedded in natural language

77 S. van Hooland, R. Verborgh, op. cit., p. 161.

78 Q. Li, *Searching for Entities: When Retrieval Meets Extraction*, dissertation, University of Pittsburgh, 2012, pp. 6–23, http://d-scholarship.pitt.edu/6227/ [access: 23/03/2016].

79 P. McNamee, J. C. Mayfield, C. D. Piatko, 'Processing Named Entities in Text', *Johns Hopkins APL Technical Digest*, vol. 30, no. 1, 2011, pp. 31–40, www.jhuapl.edu/techdigest/TD/td3001/McNamee.pdf [access: 10/04/2016].

texts, such as tables, lists, or other forms. The attributes from the tables or lists are extracted using the following rules: if the majority of the elements with the same attribute are of the same type or identified as target entities, all these elements are treated as target entities. The dictionary-based entity extraction approach is a special case of the rule-based methods. There are real-life examples of using DBpedia (the structured information version of Wikipedia built upon the Semantic Web technology stack) as dictionaries for extractions.[80] The important task in the NER technique is Entity Resolution, which involves disambiguation of the entity, and this is exactly the part of the Entity Recognition process where the additional contextualisation provided by the added values recorded in the MARC 21 658 tag fields might turn out to be useful. The repeatable MARC 21 field 658, where according to the National Library Descriptors project the content objective expressed by the covered field of knowledge ought to be recorded, might contain the following pre-defined values, covering all main areas of knowledge at a pre-defined level of granularity, namely:

- Culture and the arts
- Education
- Science and research
- Sociology and society
- Ethnology and cultural anthropology
- History
- Archaeology
- Librarianship, archive, museums
- Media and communication
- Political studies and public administration
- Law and jurisdiction
- Management and marketing
- Economy, economics and finance
- Medicine and health
- Animal care and veterinary medicine
- Architecture and construction
- Transport and logistics
- Engineering and technology
- Computers and information technologies
- Public safety and military studies
- Agriculture and forestry

80 Q. Li, op. cit.

- Biology
- Geography and Earth sciences
- Ecology and environment protection
- Chemistry
- Physics and astronomy
- Mathematics
- Philosophy and Ethics
- Literature studies
- Linguistics
- Psychology
- Religion and spirituality
- Family and social relationships
- Work, career and money
- Personal development
- Lifestyle, fashion and beauty
- Culinary art and cooking
- Hobby and leisure
- Sports and recreation
- House and garden
- Travel and tourism

To illustrate how the NER technique can benefit from this additional metadata, we can think of a very simple example where the extracted entity 'Jaguar' does not appear in the digitised text other than in the sentence 'Jaguar moves very fast.' Without too much elaborate linguistic explanation we might safely assume that neither the grammar of this sentence nor the semantics of the remaining content of the phrase can help us decide if the 'Jaguar' entity refers to the animal or to the automobile brand. We are aware that for presentation purposes this example is very trivial and that the issue could be resolved by the NER tool based on the rules from other phrases in the publication, yet it might just as well not, depending on what kind of publication we are considering. At this point it might already be obvious that with the additional information provided within the record metadata in the form of the content objective value from the 658 field, the NER technique can be supported with immediate disambiguation depending on whether it is for example 'Biology', 'Geography and Earth sciences', or 'Animal care and veterinary medicine', which would suggest the 'Jaguar' animal entity, or for example 'Engineering and technology' or 'Management and marketing', which would point on the 'Jaguar' brand. Certainly we could get further entangled in

resolving the problem if for instance the 658 field value in this case was equal to 'Ecology and environment protection', since both animal entities and cars that hugely affect the natural environment may appear in such a context. However, we must keep in mind that the 658 MARC 21 field, as a repeatable one, allows for a variety of content, and so situations when only one possible context is provided will theoretically be the less frequent ones. Since the National Library of Poland has not yet adopted the NER strategy for its huge digitised content, this aspect will certainly require further studies and is definitely not covered sufficiently within the scope of this paper, but the direct benefits which adopting the National Library Descriptors might provide for NER in the digital library should not be left unnoted.

## Conclusions

With hardware advances in instruments and data storage techniques comes the inevitable flood of scientific data that renders obsolete the traditional approaches to scientific discovery.[81] With the semantic search, the user's intentions can be accurately narrowed down in a more robust way than with the keyword search. In the semantic search, typical users (who have not had any special training) can interactively explore the data and navigate through sets of entities by utilising their knowledge of the domain. In other words, the explicit encoding of semantics by means of relations and the global referencing of entities in RDF via links or URIs are two critical enabling features that make RDF suitable for robust search and information integration across different data providers.[82] The prerequisite for this information integration and linking of various datasets is a metadata architecture where metadata fulfil certain metadata requirements. They should be simple: Metadata properties should be basic and understandable, because a simple scheme is easier to manipulate for automatic metadata generation and quality control. A simple, basic scheme can collectively support the description and discovery of heterogeneous resources, including data objects. Metadata need to be interoperable as well, as interoperability supports metadata harvesting, cross-system searching, and metadata exchange with other formats. Last but not

81 U. M. Fayyad, P. Smyth, 'Cataloging and Mining Massive Datasets for Science Data Analysis', *Journal of Computational and Graphical Statistics*, vol. 8, no. 3, 1999, pp. 589–610, www.jstor.org/stable/1390878 [access: 14/04/2016].

82 H. Haidarian Shahri, op. cit., p. 41.

least, metadata should be compatible with the Semantic Web in order to support machine processing, including automatic data synthesis, and potentially for other uses of metadata not even envisioned by their creators.[83]

## SUMMARY

This paper has provided an explanation of how the National Library Descriptors project supports the simplicity, interoperability and Semantic Web compatibility of the National Library's metadata. It has presented how the project, which initially aimed first and foremost at providing better data access by creating additional access points within the original National Library dataset of the Bibliographic Database, might also provide further advantages in combining the National Library dataset with other datasets available on the Web, as they can be relatively easily mapped to properties and classes defined in commonly used Semantic Web ontologies, and facilitate further advancements such as enabling the semantic search within the National Library's catalogue or facilitating the application of the Named Entity Recognition and Resolution techniques to the digitised context of National Digital Library.

The success of the Semantic Web will be significantly limited if content and tools are not widely shared. Much as the original World Wide Web grew from an open-source, open-content model, so must also the Semantic Web.[84] We are currently witnessing the implementation of a monumental infrastructure project, which in some ways echoes breakthrough projects of the past, such as the construction of the Roman aqueducts or the creation of the "Encyclopédie" during the Enlightenment period.[85] The National Library Descriptors project is an attempt to provide significant input into this movement that will change the way we perceive, consume, and interact with the world of data.

83 J. Greenberg, H. C. White, C. Carrier, R. Scherle, 'A Metadata Best Practice for a Scientific Data Repository', *Journal of Library Metadata*, vol. 9, no. 3–4, 2009, pp. 194–212, www.tandfonline.com/doi/abs/10.1080/19386380903405090 [access: 14/04/2016].

84 J. Hendler, op. cit., p. 521.

85 V. Mayer-Schönberger, K. Cukier, op. cit., p. 131.