IFLA WLIC 2017 WROCŁAW

Submitted on: 06/06/17

National Library of Poland Descriptors model as an attempt of opening library data for reuse

Marta Cichoń
Bibliographic Institute, National Library of Poland, Warsaw, Poland.
E-mail address: m.cichon@bn.org.pl

Abstract:

National Library of Poland introduced the Descriptors model to the structure of its authority data in the bibliographic database in order to allow better data segmentation within authority and bibliographic data and in a consequence – to enable the shift from unstructured data to structured information and to create additional links between defined entities in the National Library database thus improving possibilities for data retrieval and linking with other datasets. Data atomization and usage of standard controlled vocabularies were prerequisites for data segmentation.

In accordance to FRBR model we organized our data model basing it on the notion of entities instead of headings – they emerged from the shared pool of merged name and subject authority files. Every entity according to the entity type has a set of attributes allowing phrase based linking. The new MARC 21 fields allowed assigning additional attributes to entities, which are currently being populated in the database using the variety of methods – from manual and semi-automatic data processing based on use of regular expressions to more automatic use of matching algorithms. Changes applied to the controlled vocabulary itself were both prerequisite for data atomization and consequence of taking this approach. The achieved better data segmentation is supposed to allow populating additional facets in the faceted search of library catalog thus strongly improving user experience and possibilities of information filtering as well as to extract data with specific attributes and attribute based datasets in various data formats. To clarify and exemplify these benefits the special web based data extraction tool "data.bn.org.pl" is presented.

Furthermore, the explanation is provided of how the National Library Descriptors model supports the simplicity, interoperability and Semantic Web compatibility of the National Library's metadata.

Keywords: data reuse, metadata interoperability, data segmentation, web of data, information filtering

Background and motivation

Linking 'traditional' data sources, such as public and research data, with new sources of data, such as various web services might be a unique opportunity for complex exploration of social and cultural behaviors and newly emerging phenomena. As much as benefits of such synergy are 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.[1] Those who are willing to embrace the web of data find themselves working in a world of opportunities where they can make a real difference to the world of research. Those who are not willing to do so will discover a world of increasing change with which they are more out-of-step, finding it necessary to justify a role and practices that are not only less recognized, but in fact unnecessary.[2]

In the present world of data, the sum of information is more valuable than any of its parts and the same rule might be applied to the linked datasets. Nowadays, Internet users are already familiar with mashup web services, which are basically web sites presenting information from at least two or more sources in an innovative way, often displaying data in visual form, which makes it even more accessible for 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 multiway re-use.[3] This approach has been adopted in the National Library Descriptors project, in which we implemented the improved data model to the National Library of Poland catalogue. We took advantage of new standard MARC21 fields introduced to the MARC 21 Format mostly between years 2009-2013 along with the dawn of the RDA standard based cataloging. How exactly these changes might influence the re-use of the corresponding data will be explained further.

From unstructured to structured data

Structured data is managed by technology that allows for querying and reporting against predetermined data types and understood relationships. Structured data unlike unstructured data is anything that has an enforced composition to the atomic data types. On the other hand, unstructured data - unlike structured data - raises some distinct challenges: such data is not consumable from a semantic level without a compatible interface or application and even with a compatible technology, we cannot necessarily gain insight into the context of the information unless we can actually read it. One of the ways to approach the question of using the unstructured data, which was achievable at the National Library of Poland, is to bring the unstructured parts of the data into the structured world. If we have already identified the context and semantics of our unstructured data, we can bring this information together with our structured data.[4]

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 precisely structured fields, we need to render the meaning of the different fields (in our model shaped as attributes of model specific

[1] Internet : publiczne bazy danych i Big data, red. Grażyna Szpor, Wydawnictwo C.H. Beck, Warszawa 2014, p.55
[2] Stuart D., Factilitating access to the web of data : a guide for librarians, Facet Publishing. Beck, London 2011, pp.145-146
[3] Mayer-Schönberger, V., Cukier, K., 'Big data : rewolucja, która zmieni nasze myślenie, pracę i życie', MT Biznes, Warszawa 2014, p. 146
[4] Weglarz, G.: Two Worlds of Data – Unstructured and Structured. In: DM Review, September 2004, pp. 1-3

entities) explicit by documenting them in a schema. By structuring and atomizing 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. Through the adoption of a radically simple data model, abstraction can be made of the traditional XML and database schemas we had to use in the past to interpret and re-use data.[5]

As the notion of entities replaced headings along with the FRBR-ization we created a shared authority file for names and subjects to underline the fact that they reflected the same entities, thus allowing to get rid of data duplication in our database. The achieved better data segmentation is supposed to enable populating additional facets in the faceted search in the library catalogue and extracting records data basing on values of specific attributes. For the purpose of structuring data in the library catalogue 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. For the Authority Records the following set of MARC 21 standard fields has been added, previously unused at the National Library of Poland:[6]

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

[5] Hooland, S. van, Verborgh R., 'Linked Data for Libraries, Archives and Museum : How to clean, link and publish your metadata', London 2014, pp. 12-13
[6] https://www.loc.gov/marc/authority/, access: April 30, 2017

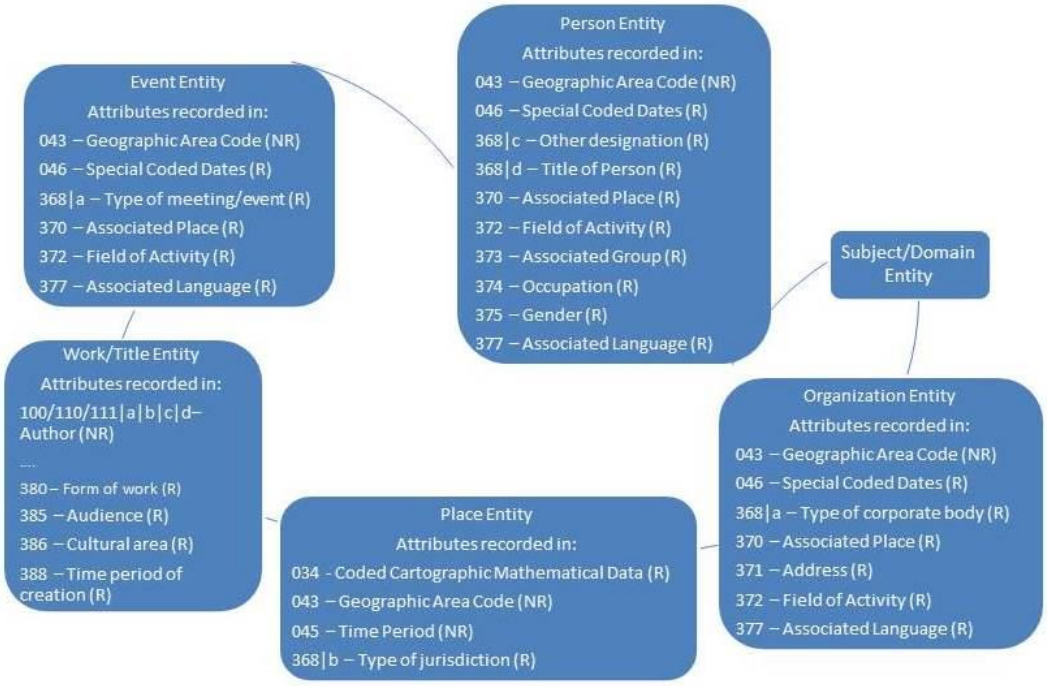
For the Bibliographic Record, the following MARC 21 fields have been added to the existing set:[7]

MARC 21 Field Tag:	
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)	

Catalogue information in the entity-relationship structure

Adding these fields to the set of already used MARC tags allowed for much more precise definition of the attributes assigned to specific headings. It aimed at creating additional relations between entities and attributes expressed by the National Library of Poland controlled vocabulary. Transformation of the used controlled vocabulary was another significant step, that was prerequisite in the implementation of the new data model, as the more atomized and simplified controlled vocabularies facilitate building relational structure. Within the described change the National Library Descriptors controlled vocabulary comprises of name headings, corporate headings, meetings, series and work uniform title as a single authority file creating a shared pool of entities which were earlier divided into name and subject authority files. This major step towards simplification aims also at clarification of relational structure as it allows to avoid the unnecessary duplicate relations and to focus on building additional relations there, where they actually provide more semantic value. Further important change within the controlled vocabulary structure was parsing the complex subject headings into atomic data elements. It was to be achieved by the fact, that for both bibliographic and authority records (in 6XX fields) the National Library will stop using subdivisions, more precisely subfields x, z, y as subfield "v" had already been not used before. All the values that were earlier recorded as general subdivisions "x", geographic subdivisions "z" and chronological subdivisions "y" are being recorded as subject in 650 fields, content objective in 658 fields or simply as geographical or chronological terms (in MARC fields 651 and 648 accordingly). The important aspect of this process is identifying the distinction between complex subject headings in which certain subdivisions represent indexed terms and where the same terms appear as the context information and therefore should be recorded in the 658 field, representing the knowledge domain, instead of 648, 650 or 651 field, representing the subject term.[8]

[7] https://www.loc.gov/marc/bibliographic/, access: April 30, 2017
[8] Żurawińska, Z., 'Deskryptory Biblioteki Narodowej w Systemie Bibliotecznym' http://www.bn.org.pl/download/document/1429787847.pdf, access: April 30, 2017



Picture 1: General entities scheme in the National Library Descriptors model

The internal architecture of the library management may make proprietary data structure more appropriate. However a bibliographic standard that is based on discrete records in a flat file structure may not be easily translated into a relational system.[9] In the implemented National Library Descriptors data model every entity according to the entity type (person, organization, event, place, work/title) has a set of attributes allowing phrase based linking and thus - developing the relational structure contributing to further advantages of data re-use.

Improved information filtering and data interoperability

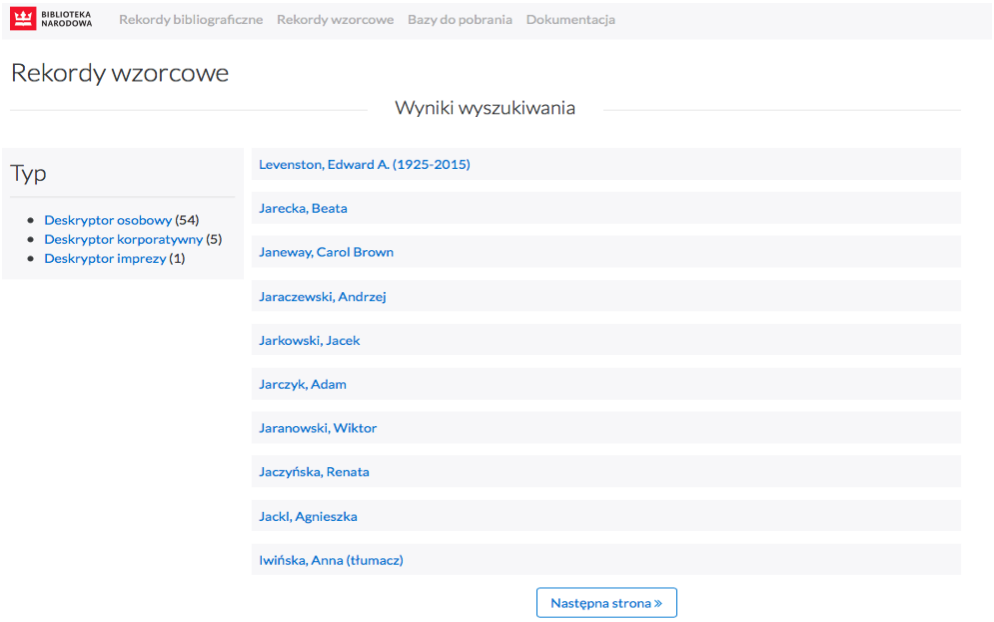
In order to present the benefits of this new data structure we introduced new web based data extraction tool available under the URL: data.bn.org.pl. It allows users to query the National Library entities pool using the information filtering possibilities that emerged from the above described data segmentation. Consequently, specific attribute based datasets can be retrieved in various formats by mean of simple queries over the http protocol. The API based service allows retrieving data in various formats - from library proprietary marc ISO2709, marcxml to xml and json standards widely used over the web.

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, which is the purpose of data translating and extracting services as the one briefly introduced above.

[9] Hayes, D., Metadata for information management and retrieval, Facet Publishing, London 2004, p.135

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.[10]

The ability to generate output in a standard format and to import records in an agreed format allows the exchange of data between systems provided that the data is capable of being used by other systems.



Picture 2: Example of the html document presenting the results of data.bn.org.pl service query: http://data.bn.org.pl/authorities?fieldOfActivity=angielski

It was earlier mentioned that the more atomized and simplified controlled vocabularies might facilitate building relational structure. This statement applies also to building relational structure of the Web by increasing the semantic interoperability. There are two contexts for metadata and interoperability: metadata as a tool to enable exchange of information between interoperating systems, and interoperability of metadata schemas themselves, which can facilitate systems' interoperability. In the area of bibliographic standards, FRBR (Functional Requirements for Bibliographic Records), to which the National Library Descriptors data model is strongly related, provided a model for bibliographic data that fostered the creation of crosswalks between schemes. Crosswalks have been published between many major metadata schemas.[11] 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 help of different vocabularies.[12]

[10] Shahri, H. H., 'On the Foundations of Data Interoperability and Semantic Search on the Web', 2011, http://drum.lib.umd.edu/handle/1903/11798, access: March 23, 2017, pp. 12-13
[11] Hayes, D., op.cit, pp. 156-161
[12] Hooland, S. van, Verborgh R., op. cit., p. 132

Future work

The rise of the Web obliged libraries and other culture curating institutions to increase the pace of their standardization 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 provide more value out of the metadata patrimony they have been building up over decades. The current hype on linked data and the Semantic Web technology underlying linked data seems to offer amazing opportunities to valorize what libraries already achieved and to facilitate the creation of new metadata.[13] This generation of Web technology is designed to improve communication between people and programmed applications 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.[14]

Although the National Library Descriptors data model initially aimed at providing better data access by creating additional access points within the original National Library of Poland Bibliographic Database, it might also provide further advantages in combining the National Library dataset with other datasets available on the Web, as the expanded relational structure can be relatively easily (in comparison to the previous data model) mapped to properties and classes defined in commonly used Semantic Web ontologies.

Ontologies connect dictionary terms with entities identified during the conceptualization 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 formalized thesauri. One of the goals of creating ontologies is to improve functioning of search and retrieval systems and information representation systems. Another example of ontologies exploit, though still related to the first one, is their implementation in the 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.[15] 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 the associated knowledge structures and inference rules.[16]

[13] Hooland, S. van, Verborgh R., op. cit., pp. 1-2
[14] Hendler, J., 'Science and the Semantic Web', January 24, 2003, www.sciencemag.org, access: January 2013, p. 520
[15] Nahotko, Marek, 'Metadane : Sposób na uporządkowanie Internetu', Kraków 2004, p. 56-57
[16] Lee, T. Berners, Hendler, J., Lassila, O., 'The Semantic Web : A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities' *Scientific American*, May 2001, p. 3

Future work includes expressing the National Library Descriptors data model as the formal ontology in order to allow the mapping and linking terms between different ontologies. By exploiting the additional semantic value of the National Library Descriptors data structure it is possible to connect to collection holdings which have been described with different vocabularies wherever links exist between the vocabularies indicating that they represent the same concept. [17] The process of mapping between ontologies includes identifying corresponding properties and classes as even if vocabularies serve different purposes, a significant overlap between the vocabularies exist. On first sight, the mapping between various vocabularies seems straightforward. For example an attempt of mapping between the National Library Descriptors properties expressing the relations between individual entities in the dataset and the standard schema.org vocabulary, which is a collaborative, community activity, founded by Google, Microsoft, Yahoo and Yandex, might include mappings such as:

MARC 21 Tag:	**schema.org vocabulary porperty**
024	schema:sameAs
034	schema:geo
046	schema:birthDate, schema:deathDate; schema:foundingDate
370	schema:birthPlace, schema:deathPlace, schema:foundingPlace, schema:location
371	schema:address
372	schema:industry
373	schema:affiliation
374	schema:jobTitle

Some of the challenges in mapping between different ontologies reside eg. in identifying the correct property restrictions or disjoint and overlapping classes which allow for better alignment process in between vocabularies, yet it is a complex endeavor. However regardless of how complex task this might be, it is a necessary next challenge currently being undertaken in the process of described changes.

Unlike more traditional areas of librarianship where there is increasing pressure to do more with less, the web of data is actually an area of growth, to which resources are increasingly being directed. Those who are willing and able to engage with the web of data will quickly realize how much there is to do, and how much they have to contribute.[18]

References

Hayes, D., Metadata for information management and retrieval, Facet Publishing, London 2004.
Hendler, J., ‘Science and the Semantic Web’, January 24, 2003, www.sciencemag.org
Hooland, S. van, Verborgh R., ‘Linked Data for Libraries, Archives and Museum : How to clean, link and publish your metadata’, London 2014.
Internet : publiczne bazy danych i Big data, red. Grażyna Szpor, Wydawnictwo C.H. Beck, Warszawa 2014.

[17] Hooland, S. van, Verborgh R., op.cit., pp. 1-2
[18] Stuart D., op. cit., p. 145

Lee, T. Berners, Hendler, J., Lassila, O., ’The Semantic Web : A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities’ Scientific American, May 2001.
Mayer-Schönberger, V., Cukier, K., ‘Big data : rewolucja, która zmieni nasze myślenie, pracę i życie’, MT Biznes, Warszawa 2014.
Nahotko, Marek, ‘Metadane : Sposób na uporządkowanie Internetu’, Kraków 2004.
Shahri , H. H., ‘On the Foundations of Data Interoperability and Semantic Search on the Web’, 2011, http://drum.lib.umd.edu/handle/1903/11798.
Stuart D., Factilitating access to the web of data : a guide for librarians, Facet Publishing. Beck, London 2011.
Weglarz, G.: Two Worlds of Data – Unstructured and Structured. In: DM Review, September 2004.
Żurawińska, Z., ‘Deskryptory Biblioteki Narodowej w Systemie Bibliotecznym’ http://www.bn.org.pl/download/document/1429787847.pdf.