(Online) ISSN 2345-0053. KNYGOTYRA. 2013. 61

POSSIBILITIES OF EXTENDING DIGITAL HUMANITIES INTO SEMANTIC WEB ENVIRONMENT: THE CASE STUDY OF CROATIAN MEDIEVAL MANUSCRIPTS AND INCUNABULA AND THEIR FRAGMENTS

Marijana Tomić | Department of Information Sciences, University of Zadar
Mirna Willer Ulica dr. Franje Tuđmana 24i, 23 000 Zadar, Croatia
E-mail: mtomic@unizd.hr, mwiller@unizd.hr

The article deals with the research on and description of medieval manuscripts and incunabula and their fragments as part of the digital humanities project in the context of the Semantic Web. It first describes the case study of Croatian medieval manuscripts and incunabula proposed as a research project by the Department of Information Sciences, University of Zadar. The focus of the project was directed at the selection of data elements required for detailed description of the mentioned material and their standardization using ontologies developed by the old and rare book community's disciplines such as bibliography, codicology, palaeography and typography. A brief introduction to the technological infrastructure of the Semantic Web and its standards follows, with a detailed description of methodology to publish a selected vocabulary in one of the metadata registry services. This section of the article concludes with the RDF graph for the partially reconstituted record of a fragment description as an example of publishing linked open data. As each of the mentioned disciplines uses its own value vocabularies or ontologies, it is suggested that mappings between their terms is designed using SKOS rules, rather than aiming at the use of one common vocabulary.

KEY WORDS: *digital humanities, medieval manuscripts, incunabula, fragments, Croatia, Semantic Web, linked open data.*

INTRODUCTION

Manuscript collections are collections of a variety of materials, defined simply as "a text or document, usually on paper or parchment, literally written by hand"[1], such as family or personal papers, diaries, letters, archive collec-

tions, etc. A special type of manuscript collections are those from the medieval period, be they codeces, maps, music material, etc., or their fragments.[2] Together with incunabula, manuscript collections are usually the most precious collections in libraries as they convey a lot of otherwise lost information on history, culture, literacy, social history and contemporary life in the medieval period. During their history, those books were ordered by numerous patrons, they were written and printed in numerous monasteries, universities and other writing or printing *officines* by numerous scribes and printers according to numerous rules connected to local or European conventions, they subsequently crossed borders, changed owners, shapes, bindings and were not just bound together but some of them were used as binding for others. Traces of history of all those people and locations are hidden in those books and fragments. By crossing national and international regions and borders, and by going from one hand to another those books hide traces of historical and cultural content which otherwise would remain lost and unknown. Therefore, researching old and rare material is of great importance for national as well as European cultural and social history.

From the point of humanities disciplines, there are several factors that are nowadays enforcing considerable changes in the manuscript and early printed books research. Among the most important is the impact of information technology on almost all aspects of research, and consequently the emersion of a new discipline called digital humanities or humanities computing[3]. This emerging field

1 BEAL, Peter. *A Dictionary of English Manuscript Terminology*. Oxford: University Press, 2009, p. 244.

2 Medieval scribes and printers often re-used the parchment and other material used for codeces and early printed books as they wanted to prevent the waste of material which was expensive and hard to get. Parchment was re-used in at least two different ways: it was scraped clean so that a new text could be written on the same sheet, or it was used as a binding material for new codeces or early prints as they were ideal material for strengthening the book's spine. In bindings, parchment was also used as endpapers or they were used as whole cover. Sometimes, all of the leaves of some manuscripts were separated from the cover and every leaf was used for covers or endpapers of other, usually printed books. There is a reason to assume that that procedure was used in printing *officines*, with manuscripts that were used as templates for printed books. After printing, a manuscript was put aside as a "manuscript waste" and was used for binding printed books. Although there are various definitions of what can be considered to be a fragment, we will use the one provided by Erich Renhart who uses the term in the sense of a part of a former book or document. More about projects on fragment description and research see in: RENHART, Erich. Manuscript fragments: the hidden library. In *Summer School in the Study of Historical Manuscripts*: proceedings. Zadar, 2013, p. 133.

3 Digital humanities was in the begging of the discipline called *humanities computing*. It is not within the range of this article to discuss that lexical shift, but more information on it can be found in a series of articles published by Patrik Svensson. See: SVENSSON, Patrik. Humanities Computing as Digital Humanities. *Digital Humanities Quarterly* 2009, 3, no. 3 [accessed 04 August 2013]. Access through Internet: <http://digitalhumanities.org/

encompasses research in humanities in and through information technology[4], which has impacted rather dramatically, opening new possibilities for research and answering questions. Owing to quite a respectable number of websites, projects and digital libraries of the medieval material, there is a vast range of material for research which enables medievalists, according to Toby Burrows, to be "at the forefront of the application of digital technologies to research in humanities"[5]. At the same time, however, T. Burrows identifies some severe limitations of those websites and services, among which he emphasizes "the lack of integration and interoperability between the many different sites", and inconsistency in the use of terminology and of descriptive standards, which causes that "researchers around the world still face major difficulties in finding, using and sharing knowledge about medieval manuscript collections"[6].

As we fully agree with T. Burrows that the solution to this problem is an international collaborative infrastructure for organizing content and interlinking knowledge[7], we see its possible realization in the technological environment of the Semantic Web and Linked Open Data. The case study of Croatian medieval manuscripts and incunabula and their fragments, described in this article, explores these technologies in the context of descriptive bibliographical and terminological issues which are prerequisite for building international collaborative infrastructure and re-use of linked open data.

A CASE STUDY OF CROATIAN MEDIEVAL MANUSCRIPTS AND INCUNABULA

The case study described in this article is derived from the project *Croatian Medieval Manuscripts and Incunabula: Towards Inclusion of Digital Humanities into the Semantic Web* which was proposed for funding to the Croatian Science Foundation by the Department of Information Sciences, University of Zadar in 2012.[8] In the focus of the project was the research of Croatian medieval manuscripts and incunabula and their fragments within digital humanities, particularly digital palaeography and codicology.

The validity of the hypothesis for the project was recognized in the state of the art in the field. Namely, there is a significant number of digital humanities projects which, alongside with analyzing, describing and editing written heritage are developing or improving digital techniques, tools and programs for investigating digitized, as well as non-digitized sources. However, we agree with Stokes who introduced to the projects at the King's College London the methodology by which traditional and digital approaches to paleographic research are combined, and who

pointed out the necessity of building some kind of collaborative environment in which remote users have access to and work on the same document(s)[9]. Such an environment raises awareness that resulting scientific research data themselves require development of an information infrastructure for storage, analysis, retrieval and re-use[10], that is, a virtual interdisciplinary research environment which interacts with technologies associated with the Semantic Web and Linked Open Data systems.

The research corpus of our project consists of a selection of digitized Croatian manuscripts and incunabula with some of their fragments written in three scripts – Glagolitic, Latin and Cyrillic, and three languages – Old Church Slavonic, Latin and Croatian type of Old Church Slavonic[11]. The research objectives based on the corpus analysis aim at defining first, the descriptive methodology of types and characteristics of scripts and languages used in medieval Croatia[12] and their mutual influence,

dhq/vol/3/3/000065/000065.html>; SVENSSON, Patrik. The landscape of Digital Humanities. *Digital Humanities Quarterly* 2010, 4, no. 1 [accessed 04 August 2013]. Access through Internet: <http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html>.

4 As defined by its "father" Roberto Busa, digital humanities are "the automation of every possible analysis of human expression (therefore, it is exquisitely a "humanistic" activity), in the widest sense of the word, from music to the theater, from design and painting to phonetics, but whose nucleus remains the discourse of written texts". See more in: BUSA, Roberto. Foreword: perspectives on the Digital Humanities. In *A companion to Digital Humanities*. Kindle Edition. Oxford: Blackwell, 2004.

5 BURROWS, Toby. Applying Semantic Web Technologies to Medieval Manuscript Research. In *Kodikologie und Paläographie im digitalen Zeitalter 2* Norderstedt, 2010, p. 118.

6 Ibidem.

7 Ibidem.

8 The project *Croatian Medieval Manuscripts and Incunabula: Towards Inclusion of Digital Humanities into the Semantic Web* was not approved.

9 PIERAZZO, Elena; Peter Stokes. Putting the text back into Context: a Codicological Approach to Manuscript Transcription. In *Kodikologie und Paläographie im digitalen Zeitalter 2 = Codicology and Palaeography in the Digital Age 2*. Ed. by Franz Fischer, Christiane Fritze, Georg Vogeler. Norderstedt: BoD, 2010, p. 397–430; STOKES, Peter. Palaeography and image processing: some solutions and problems. *Digital Medievalist*, 2007/2008, no 3, [accessed 04 August 2013]. Access through Internet: <www.digitalmedievalist.org/journal/3/stokes/>.

10 BORGMAN, Christine L. *Scholarship in the Digital Age*. Cambridge, 2007.

11 Croatian medieval period is characterized by close contact and interweaving of abovementioned three scripts and three languages. More about this see in: HERCIGONJA, Eduard. *Tropismena i trojezična kultura hrvatskoga srednjovjekovlja*. Zagreb: Matica hrvatska, 1994; PELC, Milan. *Pismo – knjiga – slika: uvod u povijest informacijske kulture*. Zagreb, 2002.

12 Some of the projects aiming to describe Crotian Glagolitic script using IT are conducted in the Institute for Slavonic Studies of the University of Vienna. See more in: DIEM, Markus; Robert Sablatnig; Melanie Gau; Heinz Miklas. Recognizing Degraded Handwritten Characters. In *Kodikologie und Paläographie im digitalen Zeitalter 2*. Norderstedt, 2010, p. 295–306.

and second, material characteristics of manuscripts and early printed books and their fragments, specifically their codicological features, page layout and graphitic features[13] which are used for text ubication and datation. Defining features of script and the abovementioned material characteristics, and activities on systematizing terminology in the field of codicology and palaeography is an important prerequisite for standardization of the description of manuscripts and incunabula, and, as to their fragments, for enabling virtual integration of the same codicological unit. The third objective is aimed at implementing methodologies for publishing vocabularies in the field of codicology and palaeography, together with their mutual alignments, as well as their alignments with library-based standards in order to publish metadata content as linked open data within the Semantic Web of textual heritage.

VOCABULARIES FOR DESCRIBING MANUSCRIPTS AND INCUNABULA AND THEIR FRAGMENTS

Due to the fact that bibliographic description of manuscripts is not prescribed as precisely as that for other library material, it is mainly done by humanists, for instance codicologists, palaeographers or bibliographers of old and rare material in their own idiosyncratic manner. Description of fragments is even less standardized, although it would be required in the context of collaboration among institutions which, enabled by digital environment opens up possibilities of matching and reunion of scattered fragments. In both cases, the need for highly standardized vocabularies is recognized.

According to Peter Beal, when describing a manuscript or an incunabula, the cataloguer usually describes their physical form or nature of their contents or function they perform (psalter, bible, breviary, patent, history book), physical materials associated with them (paper, ink, vellum), their other physical features (size, textual layout, decoration, collation, bindings, etc.), their condition, writing instruments, writing surfaces, manuscript containers, as well as scribes and other functionaries responsible for document production, notable custodians and repositories of manuscript, scripts and handwriting, lettering and paleographical features[14]. However, there is no consistency in using all of the mentioned elements. Information depth of a record in a manuscript catalogue is not prescribed, and depends on the cataloguer's decision. That is why we are faced with an unbalanced content in manuscript and incunabula catalogues. While some records are very detailed and have almost all of those elements, some have just a few of them. Various national schools developed at different periods their rules for describing or cataloguing manuscripts but none of them are accepted as an international standard.

Based on Peter Beal's compilation of elements for a full and detailed description of manuscripts, we propose that every description contains, where applicable, the following:

- title of a work (with first line if a poem, or incipit if a medieval text or a song)
- genre or type of a document
- contents
- authorship of the text(s) (identified, attributed, or ascribed)
- material (paper, parchment, etc.)
- watermark(s) (in paper)
- identity of handwriting
- script(s) employed
- special or characteristic features of handwriting
- leaf size or format
- foliation or pagination
- total number of leaves, pages, or membranes
- layout
- decoration, illustration, or illumination
- collation
- binding
- evidence of provenance
- any other features of interest (e.g., annotations, side-notes, glosses, index)
- present location and press-mark

Fragment studies, although rare, are considered complementary to manuscript studies as fragments are usually the only remains of otherwise lost manuscripts and incunabula[15]. There are three main goals of fragment studies that can be

13 Graphetic features are defined in grapholinguistics, a branch of palaeography, as features inherent to the script which do not have a direct parallel in pronunciation, but present a visual surplus that participate in an adequate visual presentation of written material. More about grapholinguistics and graphetic features see in: ŽAGAR, Mateo. *Grafolingvistika srednjovjekovnih tekstova*. Zagreb, 2007, p. 11–13; *Lexicon der germanistischen Linguistik*. Tübingen, 1973, p. 105–110, 118–132.

14 BEAL, Peter. *A Dictionary of English Manuscript Terminology*. Oxford, 2009, p. IX.

15 More about Austrian project on fragments research and description see in: RENHART, E. Op. cit., p. 133; One of the pioneer projects on Glagolitic fragments in Croatia is a the one on Glagolitic fragments found in the libraries of Capuchin monasteries, conducted by Anica Vlašić-Anić from Old Church Slavonic Institute, Zagreb, Croatia. Fragments found within that project which are describe from the paleographic, linguistic and cultural history point of view, and its datation and authorization, and, when possible, reunion of scattered fragments belonging to the same codicological unit (same codex), are based on precise and detailed paleographical research and description. Although this is a worthy and extremely important project for Croatian cultural and written heritage, and

identified in the literature and projects: (1) *describing, digitizing and presenting* material in catalogues and other data bases; (2) *researching* old and rare material based on codicological and palaeographic analysis; (3) *matching fragments of the same codicological unit* scattered in different institutions, even in different countries or continents, based on metadata in catalogue records of fragments and/or digital images of fragments as part of online collections.

All of the elements enumerated by P. Beal are recommended for fragment description too. Nevertheless, in order to match scattered fragments and re-unite the codicological unit, additional elements are needed:

- language
- content
- detailed information on physical appearance:
  - … format (format of the current state of the fragment, proposed measures calculated according to medieval layout laws)
  - … page layout (justification, column width and height, intercolumn, measurement of text block and of the margins, number of lines, interlinear space, etc.)
  - … script (description of palegraphic features specific for a scribe or a scriptorium that may have a crucial role in matching scattered fragments of the same codicological unit – type of script, height of letters specific for the

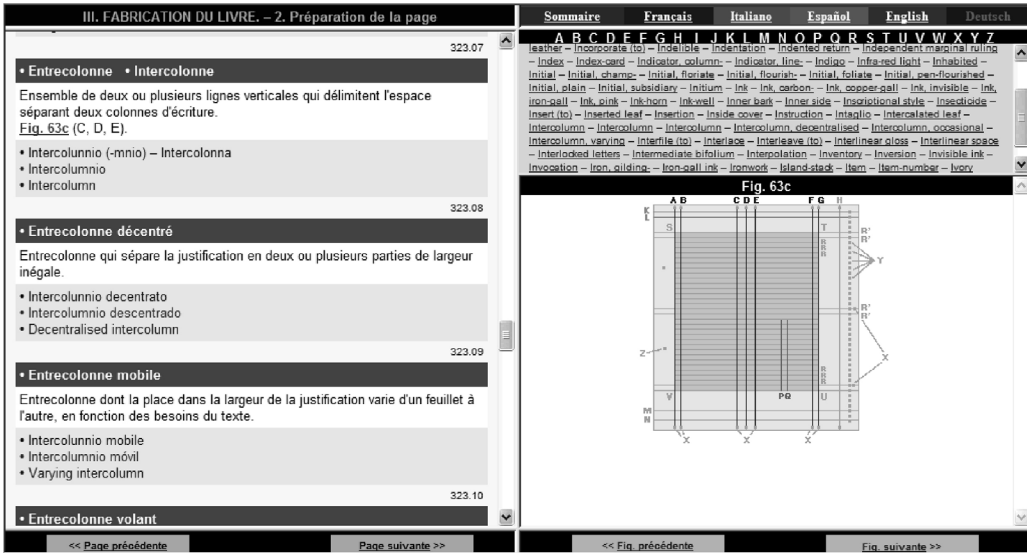


FIGURE 1. Term with its parallel language forms and its definition in Muzerele's *Vocabulaire codicologique* – entrecolonne (323.07)

script used in the fragment, usage of abbreviations and ligatures, usage of punctuation)
- ... usage of catchwords and their form and position, usage of foliaton or pagination and their form and position
- ... damages (including its provenance and position)
- ... manufacturing clues (codicological punctuation with its description and measures, rulling traces, sewing or gluing traces, etc.)
- ... illuminations and illustrations (type of initials, group of initials, type of the border illuminations, height of initials, colours in use, usage of gold or other metals, etc.)

Obstacles that occurred when trying to define elements proposed for description and matching fragments, and to find adequate terminology were: (1) enumerated features are inherent to various disciplines each of which has its own terminology, such as paleography, codicology, bibliography, typography; (2) across various disciplines connected with manuscripts and early prints the same term is used for different elements, or, which is more often, different terms or descriptive expressions are used for the same element. To overcome these obstacles that are preventing identification and re-use of descriptive metadata with the goal to matching fragments, it is necessary to identify terminological vocabularies that are used by researchers, and to design mappings or alignments between their terms within the proposed information and technological environment of the Semantic Web and linked open data.

Vocabularies mostly consulted by codicology and bibliographic community and chosen as those that can be adequately used for our purposes are those of Denis Muzerele[16], Marilena Maniaci[17], Peter Beal[18], and Christine Jakobi-Mirwald[19]. The test we performed on their applicability to detailed manuscript and incunabula

all of the fragments are described and researched in much details, it lacks the research data available for re-use by other researchers that would allow matching of those fragments with other Glagolitic fragments found in other Croatian and European institutions. More about the project in: VLAŠIĆ-ANIĆ, Anica. Novootkriveni glagoljski fragmenti u riječkoj kapucinskoj knjižnici. In *Sprache und Leben der frühmittelalterlichen Slaven*. Frankfurt am Main, 2010, p. 95–124; VLAŠIĆ-ANIĆ, Anica. Fragments of parchment codices on the covers of old books. In *Summer School in the Study of Historical Manuscripts*. Zadar, 2013, p. 145–172.

16 MUZERELE, Denis. *Vocabulaire codicologique*: répertoire méthodique des termes français relatifs aux manuscrits. Paris, 1985 [accessed 4 August 2013]. Access through Internet: <http://vocabulaire.irht.cnrs.fr/pages/vocab2.htm>.

17 MANIACI, Marilena. *Terminologia del libro manuscritto*. Roma, 1997.

18 BEAL, Peter. *A Dictionary of English Manuscript Terminology*. Oxford, 2009.

19 JAKOBI-MIRWALD, Christine. *Buchmalerei*: ihre Terminologie in der Kunstgeschichte. Berlin, 1997.

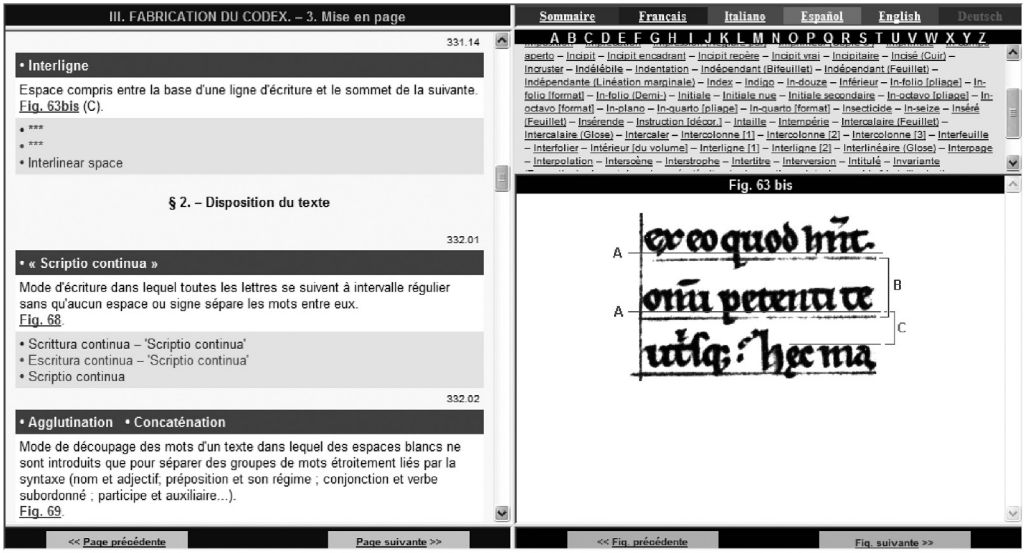


FIGURE 2. Term with its English parallel form and definition in *Vocabulaire codicologieque* – interligne (331.14)

and their fragments description showed that all of them fully complied with the given requirement.

The requirement that was tested was description of two features connected with page layout which are important for manuscript and incunabula description and fragment matching: intercolumn and interlinear space. *Intercolumn* presents the space (width) between the two columns of the text block. Although it is extremely rarely used in descriptions, it is an important part of the page layout, especially in the context of fragment matching. When it is present in descriptions, variant terms are used such as *entrecolonne* and *intercolonne* in French, *space between columns* and *intercolumn* in English, or some other descriptive forms. Figure 1 shows the French term *entrecolonne* with its parallel Italian, Spanish and English forms as defined in Muzerele's *Vocabulaire codicologique*.

The same problem is identified with the French term *interligne*; parallel terms used in English are *interlinear space* and *space between lines*, while it is found to be also referred to in some other descriptive forms. In *Vocabulaire codicologique* it is defined as a space between lines (height) (Figure 2).

Figure 3 shows four fragments of the same book, *Način za dobro umrit*, found scattered in the library of Franciscan monastery on the island Preko near Zadar,

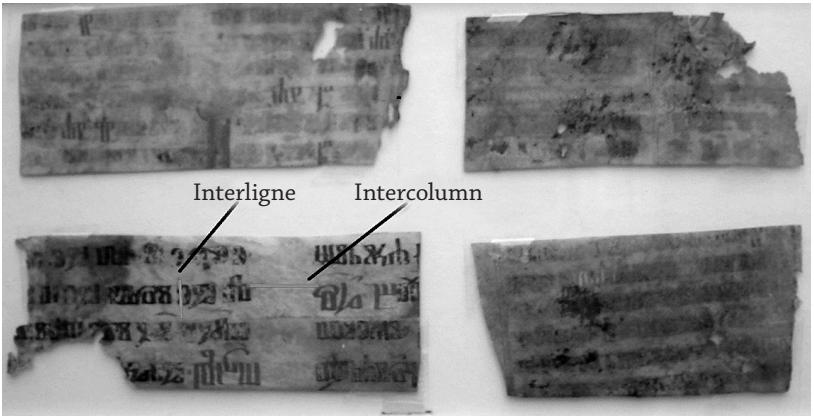


FIGURE 3. Fragments of the book *Način za dobro umrit*, Franciscan monastery on the island Preko near Zadar, Croatia

Croatia. The relevant factor in matching these fragments is the measurement of intercolumn and interligne.

SEMANTIC WEB AND PUBLISHING VOCABULARIES AS LINKED OPEN DATA

Semantic Web as the new technological environment is important to be considered from basically two points relevant to our topic: the intention of any research is to be published in order to be consumed and used for further research and/or implementation, and the publication of a research should be published in a "language" – a vocabulary understood by as many potential users as possible. That means that vocabularies that are being used in the present technology – either in print or an online version for consumption by *humans*, have to be transferred onto the new technological platform in which they will be represented for *machine* inference or "reasoning", not for humans to consume. The idea of the Semantic Web is to enable "navigable", shared space as a new genre of communication in which self-describing documents exist, based on URIs – unique identifiers of "things" on the Web'[20].

A brief introduction to the technological infrastructure of the Semantic Web and its standards will be presented first, followed by the description of one of the metadata registry services in which the vocabularies mentioned in the previous

20 BERNERS-LEE, Tim. Semantic Web – XML-2000: Philosophy – „Web" [accessed 4 August 2013]. Access through Internet: <http://www.w3.org/2000/Talks/1206-xml2k-tbl/Overview.html>.

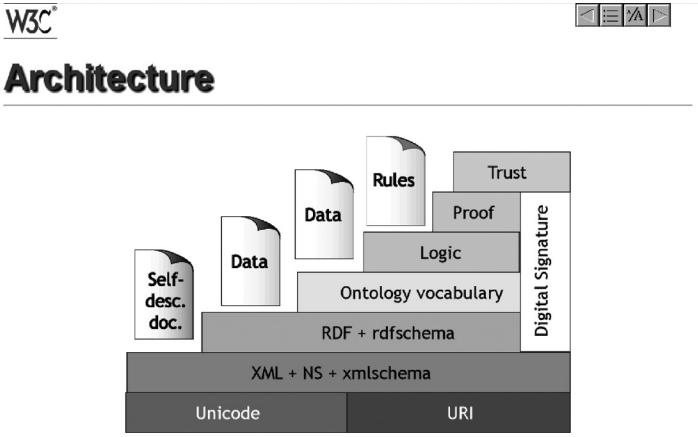


FIGURE 4. Tim Berners-Lee's *layer-cake diagram of the Semantic Web*

section can be published. The intention of this section is not to give a comprehensive description of the technological environment and state of the art in this field[21], but to alert the rare books and manuscripts community to consider publishing its valuable assets for use and re-use in the environment of linked open data promoting authorized and validated content sharing.

RULES FOR PUBLISHING VOCABULARIES. The author of the World Wide Web Tim Berners-Lee presented the architecture of the Semantic Web as a *layer-cake diagram of the Semantic Web* (Figure 4)[22]. This seems to be the simplest depiction of technological requirements for publishing content as linked open data.

The architecture consists of seven layers of which the first of the two lower ones which describe technological requirements is the one for characters encoding: UNICODE, and URI: Uniform Resource Identifier to uniquely identify a „thing" or anything one wants to „talk" about or publish as linked open data. The second layer consists of the xml schema which ensures the composition of a „message"– the syntax rules. On the top of these is the layer of the so called "rules" for self-describing documents – RDF: Resource Description Framework[23] which is a de facto standard model for data interchange on the Semantic Web. The standard supports simple, single metadata statements known as triples: *subject – predicate – object*. RDF requires the subject and predicate of a triple to be URIs, while the object can be either a URI or a literal string, that is, a content value for humans to read. If an object is a URI, it means that it becomes a subject of another triple, which object can again be a URI or a literal. This is the main idea behind the linked data concept.

The next layer of the Semantic Web architecture deals with the ontology vocabularies, that is, the rules for expressing vocabularies. As already mentioned, in the context of the Semantic Web terms used for, in our case codicological description, should be registered and defined first in order that machines can make inference about (meta)data they encode.. In this instance, it is important to decide what kind of content or data is intended to be expressed as linked data: is the content a literal, such as the title or first words of a fragment, or a term or concept from a value vocabulary, such as a language or script code, or an authorized subject or descriptive term. In the latter case, the W3C: World Wide Web Consortium, the organization that is responsible for web standards and recommendations, published specification SKOS: Simple Knowledge Organization System which 'provides a model for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, folksonomies and other similar types of controlled vocabulary'[24].

Implementation of these basic rules can be exemplified by the process of describing one of the fragments of the book which were found scattered in the library of Franciscan monastery at island Preko near Zadar shown in Figure 3, specifically the two already mentioned features of the page layout – intercolumn and interligne. Our task is first to decide what kind of vocabulary these terms belong to. The term can be treated as a *value vocabulary*, that is an authorized term that functions as an *object of a triple* and therefore is represented by a URI, which itself becomes a subject of another triple in which the value of the term is displayed (by use of SKOS rules) as a literal to the human user in the form and language appropriate to his or her needs. The term can be considered also to be a descriptive one, that is, it can describe the page layout by stating that the *interligne* (interlinear space) of the fragment is 0,4 cm width; according to the RDF rules it is considered an *element* which functions as a *predicate of a triple*.

In order to describe the process of registering a vocabulary and subsequently publishing data as linked open data, we will take as example a description of a fragment that can include data shown in Table 1, which, for the need of being expressed as a set of statements of an RDF triple is laid out in Table 2.

21 For detailed description of the new technological environment relevant to the library and information community see: WILLER, Mirna; DUNSIRE, Gordon. *Bibliographic Information Oganization in the Semantic Web*. Cambridge, 2013.

22 BERNERS-LEE, Tim. Semantic Web – XML-2000: Philosophy – „Web" [accessed 4 August 2013]. Access through Internet: <http://www.w3.org/2000/Talks/1206-xml2k-tbl/Overview.html>.

23 *Resource description Framework*: RDF [accessed 04 August 2013]. Access through Internet: <http://www.w3.org/RDF/>.

24 W3C. SKOS Simple Knowledge Organization System Primer, W3C Working Group Note 18, August 2009 [accessed 04 August 2013]. Access through Internet: <www.w3.org/TR/skos-primer>.

TABLE 1. Attributes and their values for the description of a fragment (selection)

Field/Attribute	Value
Record Identifier	160884
Title of a work (or first line…)	Način za dobro umrit
Language	Old Church Slavonic
Script	Glagolitic
Material	Paper
Content form	Text
Entrecolonne	0,3 cm
Interligne	0,4 cm

TABLE 2. Description of a fragment formatted as a set of statements

Record Identifier	Attribute	Value
160884	(has) title proper	Način za dobro umrit
160884	(has) language	Old Church Slavonic
160884	(has) script	Glagolitic
160884	(has) material	Paper
160884	(has) content form	Text
160884	(has) entrecolonne	0,3 cm
160884	(has) interligne	0,4 cm

A full description of the methodology to produce linked data triples is described as a 9 stage process by Dunsire[25], but here we will discuss only the question of how to represent terms from Muzerele' codicological vocabulary which was chosen for as an example vocabulary. Following the above analysis, terms can be defined as descriptive, and therefore should be represented as element sets. However, as the vocabulary is hierarchically structured, and there may be a need to design mappings with other relevant vocabularies, terms may be registered also as value vocabularies.

METADATA REGISTRY SERVICE. In order to publish a vocabulary, it is necessary to represent its terms in the required language, that is, each term has to be identified with a URI within its vocabulary or terminological list. In order to do that, it is necessary to choose a maintenance service that ensures open access and stability, is sustainable and expandable, as well as the one that enables interoperability and conversion of element sets and vocabularies into other registry environment in case of the need for migration. The service that IFLA: International Federation of Library Associations and Institutions has chosen for publishing its standards and models is OMR: Open Metadata Registry[26] maintained by the MMA: Metadata Management Associates, a non-profit organization.

The first step in registering a vocabulary, in our case

Muzerele, Denis. *Vocabulaire codicologique : répertoire méthodique des termes français relatifs aux manuscrits*. Paris: Editions CEMI, 1985. Available at: http://vocabulaire.irht.cnrs.fr/pages/vocab2.htm,

is to register its URI which can be composed of the original URL to which we can add "terms", so it can be ***http://vocabulaire.irht.cnrs.fr/pages/vocab2/terms/***,

or better, a new web base domain may be registered as ***http://muzerele.info/vocabulaire/terms/***.

The base domain .info is recommended for such use. In the case we decide to register Muzerele's vocabulary as an element set, the URI would be: ***http://muzerele.info/vocabulaire/elements/***.

It should be noted that it is not "un-legal" to register the whole vocabulary as a value vocabulary with respect to its hierarchical structure, and as an element set.

The methodology to register vocabulary terms will be described first (Figure 5), followed by the one for registering element sets (Figure 6). Depending on the requirement to represent the whole hierarchical structure within which the term belongs, or only the position of a term within the structure, each level may be registered by a URI. A collection of all the URIs using the same base domain is called a "namespace".

The chosen term – "Interligne" is situated within the following hierarchical structure:

3. LA FABRICATION DU LIVRE
33. Mise en page
331. Utilisation de la page
331.01 Mise en page
.....
331.14 Interligne

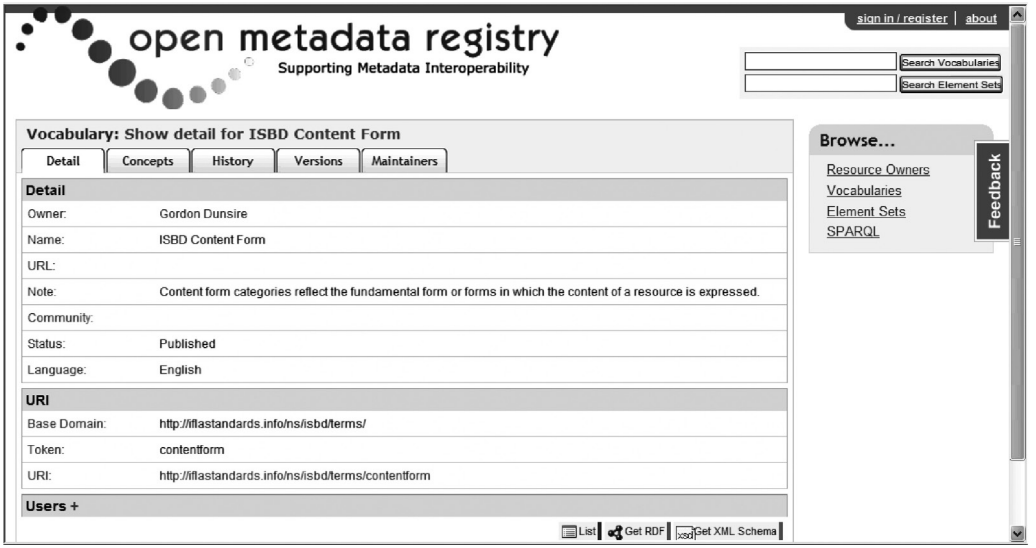
Additionally, it has to be decided whether to express the term or element by literals or numbers; as Muzerele's vocabulary uses numbers to denote a term within the hierarchy, and also defines parallel terms of this multilingual vocabulary, it is easy to decide to use numbers rather than the French original term. Thus, the term "Interligne" can be represented by a URI in the following way: *http://muzerele.info/vocabulaire/terms/fabrication/T33114*.

Such a term can be registered in the OMR following the template of the registered ISBD term "Content Form". First, the category of the terms is registered –

25 DUNSIRE, Gordon. Linked data for manuscripts in the Semantic Web. In *Summer School in the Study of Historical Manuscripts*. Zadar, 2013, p. 199–218 [accessed 4 August 2013]. Access (to preprint) through Internet: <http://www.gordondunsire.com/pubs/docs/LinkedDataForManuscripts.pdf>.

26 *Open Metadata Registry* [accessed 4 August 2013]. Access through Internet: <http://metadataregistry.org/>.

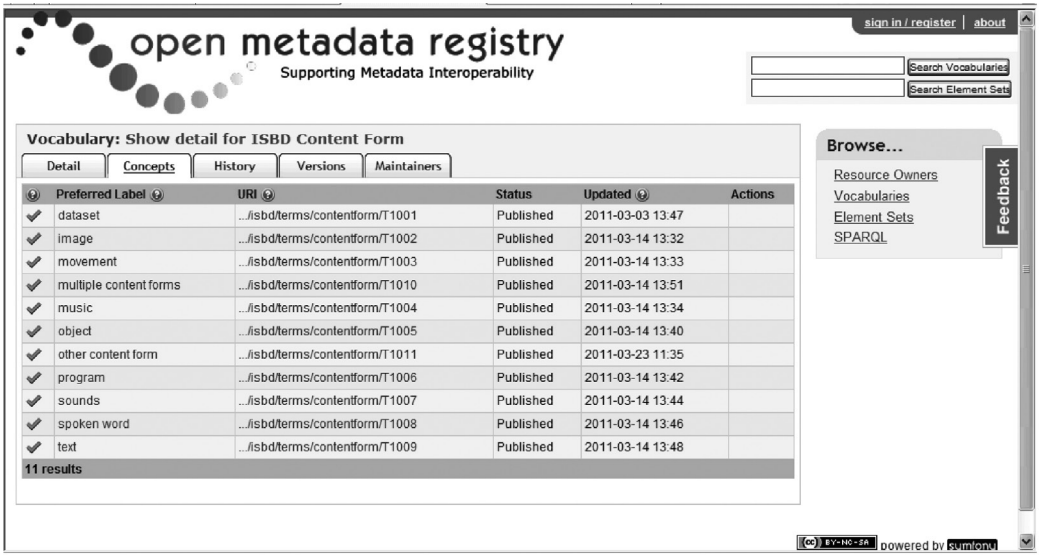
"Fabrication" (Figure 5.1), then terms that are on the lowest level as concepts – "Interligne", etc. (Figure 5.2), and finally the detail for "Interligne" itself (Figure 5.3). The URI can be constructed in a similar way for other terms and their categories: for the category 1. Les supports de l'écriture the URI may be *http://muzerele.info/vocabulaire/terms/support/*.



as follows:

Vocabulary: Show detail for Muzerele Vocabulaire codicologigue Fabrication				
Detail	Concepts	History	Versions	Maintainers
Owner:	irht.cnrs.fr			
Name:	Muzerele Vocabulaire codicologigue Fabrication			
URL:	http://vocabulaire.irht.cnrs.fr/pages/vocab2.htm			
Note:	Fabrication du livre			
Community:				
Status:	New-Proposed			
Language:	French			
URI				
Base Domain:	http://muzerele.info/vocabulaire/terms/			
Token:	Fabrication			
URI:	http://muzerele.info/vocabulaire/terms/fabrication/			

FIGURE 5.1. Namespace for the categories ISBD Content Form[27] and Muzerele Vocabulaire codicologigue Fabrication



as follows:

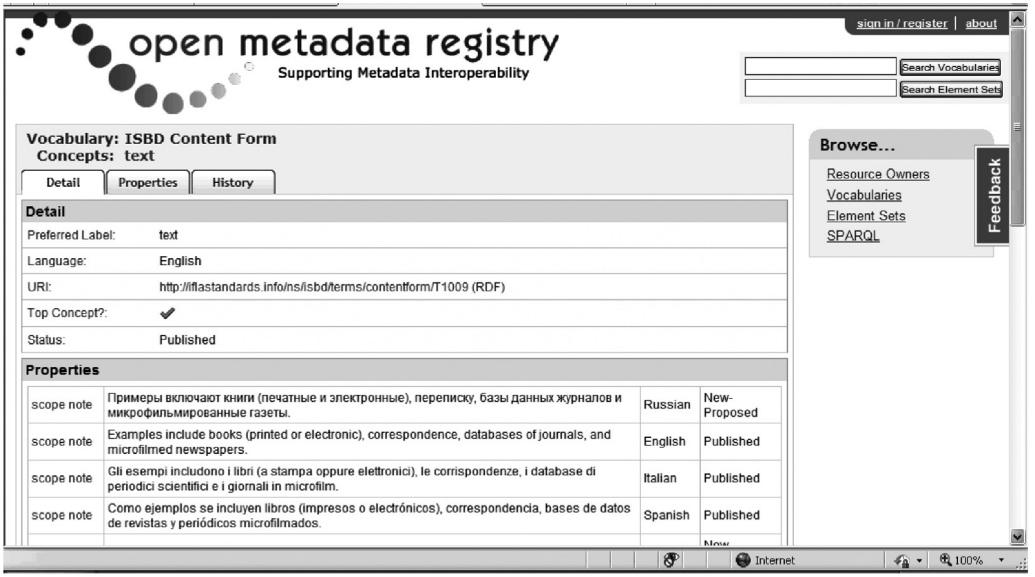
Vocabulary: Show detail for Muzerele Vocabulaire codicologigue Fabrication

Detail	**Concepts**	History		Versions	Maint...
?	**Preferred Label**	**URI**	**Status**	**Updated**	**Actions**
✔	mise en page	.../terms/fabrication/T33101	New-Proposed		
✔	opisthographe	.../terms/fabrication/T33102	New-Proposed		
✔	anopisthog-raphe	...terms/fabrication/T33103	New-Proposed		
✔	copie imposee	.../terms/fabrication/T33104	New-Proposed		
	[etc.]				
✔	interligne	.../terms/fabrication/T33114	New-Proposed		
14 results					

FIGURE 5.2. Concepts within the category ISBD Content Form[28] and Muzerele Vocabulaire codicologigue Fabrication with their namespaces that correspond to the number identifier within the ISBD Content Form and *Vocabulaire codicologigue* respectively

27 Access through Internet: <http://metadataregistry.org/vocabulary/show/id/113.html>.

28 Access through Internet: <http://metadataregistry.org/concept/list/vocabulary_id/113.html>.

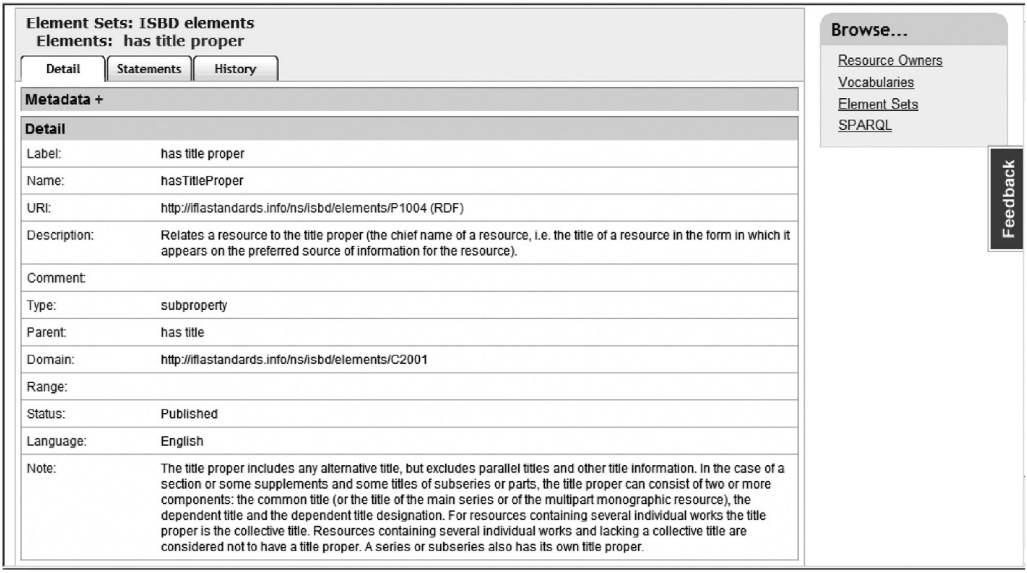


as follows:

Vocabulary: Muzerele Vocabulaire codicologigue Fabrication **Concept**: interligne			
Detail	Properties	History	
Detail			
Preferred Label:	Interligne		
Language:	French		
URI:	http://muzerele.info/vocabulaire/terms/fabrication/T33114 (RDF)		
Top Concept?	✔		
Status:	New-Proposed		
Properties			
scope note	*****	English	New-Proposed
scope note	*****	French	New-Proposed
preferred label	interlinear space	English	New-Proposed
preferred label	interligne	French	New-Proposed
definition	****	English	New-Proposed
definition	Espace compris entre la base d'une ligne d'écriture et le sommet de la suivante	French	New-Proposed

FIGURE 5.3. Detail for the ISBD concept "text" and "interligne" within the category Muzerele Vocabulaire codicologigue Fabrication with its namespace that correspond to the number identifier within the *Vocabulaire codicologigue*, and properties such as Scope note, Preferred label and Definition in different languages; scope note is lacking in both languages, while definition in English[29]

Registration of the Muzerele's *Vocabulaire codicologigue*'s element "interligne" can follow the pattern of the registration of any ISBD property, for example "has title proper". Apart from defining the URI, each property may declare its *domain*, that is, the class (subject) to which it belongs, and *range*, that is the class of the object or the value of the property. In our case the domain may be declared as *fragment* – C2001, which can be defined as a sub-class of ISBD class *resource* – C2001. The range may remain undeclared as the value of the property is expected to be literal (0.4 cm). Furthermore, we may define this element as well as all other elements of the same type as sub-property to the element *has page layout*. This element can serve the function of display of page layout characteristics as an aggregated element of all the values of its sub-properties.



Element Sets: ISBD elements
Elements: has title proper

Detail | Statements | History

Browse...
Resource Owners
Vocabularies
Element Sets
SPARQL

Feedback

Metadata +

Detail

Label:	has title proper
Name:	hasTitleProper
URI:	http://iflastandards.info/ns/isbd/elements/P1004 (RDF)
Description:	Relates a resource to the title proper (the chief name of a resource, i.e. the title of a resource in the form in which it appears on the preferred source of information for the resource).
Comment:	
Type:	subproperty
Parent:	has title
Domain:	http://iflastandards.info/ns/isbd/elements/C2001
Range:	
Status:	Published
Language:	English
Note:	The title proper includes any alternative title, but excludes parallel titles and other title information. In the case of a section or some supplements and some titles of subseries or parts, the title proper can consist of two or more components: the common title (or the title of the main series or of the multipart monographic resource), the dependent title and the dependent title designation. For resources containing several individual works the title proper is the collective title. Resources containing several individual works and lacking a collective title are considered not to have a title proper. A series or subseries also has its own title proper.

as follows:

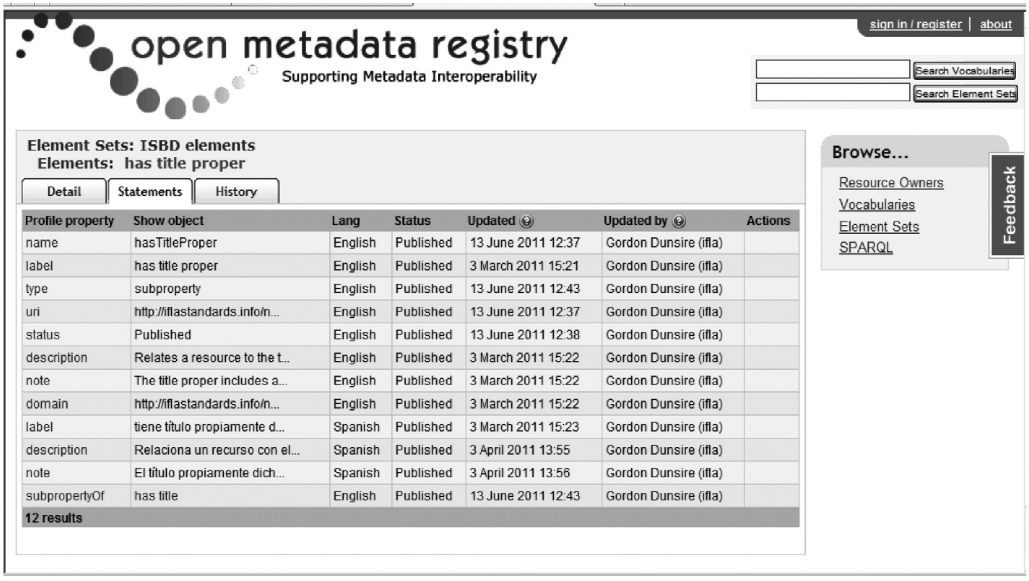
Element Sets: Muzerele Vocabulaire codicologigue Elements Elements: has interligne			
Detail	Statements	History	
Metadata			

29 For the full view of ISBD Content form text in OMR, see: <http://metadataregistry.org/concept/show/id/1208.html>.

Detail	
Label:	has interligne
Name:	hasInterligne
URI:	http://muzerele.info/vocabulaire/elements/T33114 (RDF)
Description:	Espace compris entre la base d'une ligne d'écriture et le sommet de la suivante
Comment:	
Type:	Subproperty
Parent:	has page layout
Domain:	http://muzerele.info/vocabulaire/elements/C2001
Range:	
Status:	New-Proposed
Language:	French
Note:	

FIGURE 6.1. Namespace for the element ISBD has title proper[30] and Muzerele Vocabulaire codicologigue has interligne

As Muzerele's *Vocabulaire codicologigue* is multilingual, so the elements should be represented in the languages other than French too. Figure 6.2 shows the methodology of registering the label, description and note in one of those languages.



as follows:

Element Sets: Muzerele Vocabulaire codicologigue Elements Elements: has interligne						
Detail	Statements	History				
Profile property	**Show object**	**Lang**	**Status**	**Updated**	**Updated by**	**Actions**
name	hasInterligne	French				
label	has interligne	French				
type	subproperty	French				
uri	http://muzerele....	French				
status	New-Proposed	French				
description	Espace compris...	French				
note						
domain	http://muzerele....	French				
label	has interlinear space	English				
description	******	English				
note	******	English				
subproperty of	has page layout					

FIGURE 6.2. Statements for the element ISBD has title proper[31] and *Muzerele Vocabulaire codicologigue* has interligne in two languages; description and note are lacking in English; status is New-Proposed

In order to publish description as linked open data, we have to prepare a valid RDF statement: *subject – predicate/property – object/value*. The values of the statements corresponding to the example record in Table 2 are replaced with URIs where available, as shown in Figure 7.1. It can be noticed that the properties belong to different namespaces, namely, ISBD for element *title proper* and vocabulary *text*, Dublin Core term for *language*, RDA: Resource Description and Access for graphic term *paper*, and Muzerele's elements for *interligne* and *entrecolonne*. This is legal, as long as domains of the properties refer to the same class[32]. Figure 7.2 shows the RDF graph for the same partially reconstituted record.

30 Access through Internet: <http://metadataregistry.org/schemaprop/show/id/1945.html>.

31 Access through Internet: <http://metadataregistry.org/schemapropel/list/schema_property_id/1945.html>.

32 There are certain constraints in this respect, but here we will not deal with that issue. For more information about constrained and unconstrained elements, see WILLER, Mirna; DUNSIRE, Gordon. *Bibliographic Information Oganization in the Semantic Web*. Cambridge, 2013 (October).

Subject URI	Attribute property URI	Value
mxl:160884	isbd:P1014	Način za dobro umrit
mxl:160884	dct:language	[Old Church Slavonic]*
mxl:160884	[(has script)]*	[Glagolytic]*
mxl:160884	rdaGr1:baseMaterial	rdabm:1011
mxl:160884	isbd:P1018	isbdcf:T1009
mxl:160884	muzerele:T32307	0,3 cm
mxl:160884	muzerele:T33114	0,4 cm

FIGURE 7.1. Values of the statements corresponding to the example record in Table 2 replaced with URIs where available [* Not yet registered as a property and values]

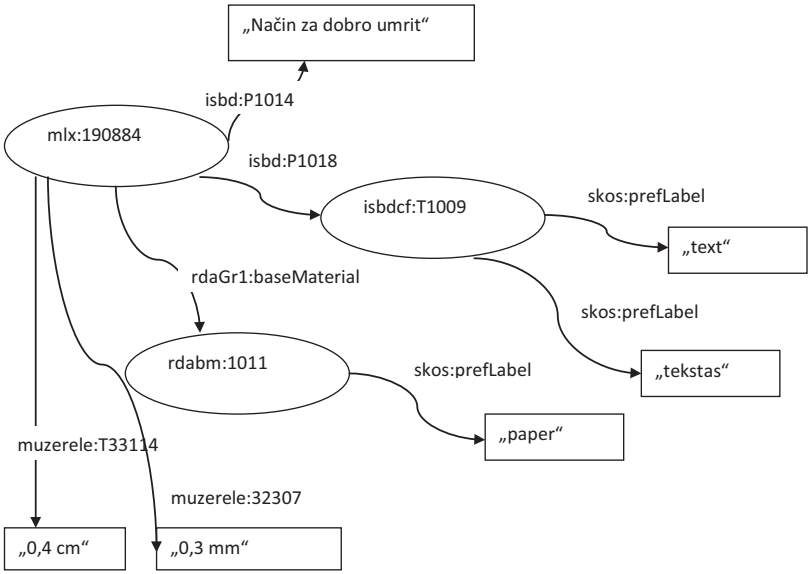


FIGURE 7.2. RDF graph for the partially reconstituted example record in Table 2

Conclusion

Manuscripts and incunabula and its fragments are of great importance for national and international cultural history and as such deserve the full attention to be researched and described. Availability of digitized material and scientific research data present a great opportunity for their remote re-use, but the obstacles to that are seen in the lack of collaborative infrastructures or integration and/or interoperability between different sites, and inconsistency in the use

of terminology and descriptive standards. The latter obstacle is seen as the major cause that researchers around the world still face difficulties in finding, using and sharing knowledge about medieval manuscript collections.

This article describes the possible solution to this problem in the implementation of technologies of the Semantic Web and the Linked Open Data. The methodology for publishing value vocabulary and element sets in a metadata register service is described on the example of Muzerele's *Vocabulaire codicologique*. An example record which uses the registered vocabulary for the description of a fragment is presented in a table and RDF graph forms to explain the methodology for publishing linked data. As the Muzerele's vocabulary is one of the available and widely used vocabularies, it is recommended to design mappings or alignments between them using SKOS rules within the same metadata registry service.

Literature

1. BEAL, Peter. *A Dictionary of English Manuscript Terminology*. Oxford: University Press, 2009.
2. BERNERS-LEE, Tim. Semantic Web – XML2000: Philosophy – „Web" [accessed 4 August 2013]. Access through Internet: http://www.w3.org/2000/Talks/1206-xml2k-tbl/Overview.html.
3. BORGMAN, Christine L. *Scholarship in the Digital Age*: information, infrastructure and the Internet. Cambridge: MIT Press, 2007.
4. BURROWS, Toby. Applying Semantic Web Technologies to Medieval Manuscript Research. In *Kodikologie und Paläographie im digitalen Zeitalter 2 = Codicology and Palaeography in the Digital Age 2.* Ed. by Franz Fischer, Christiane Fritze, Georg Vogeler. Norderstedt: BoD, 2010, p. 117–131.
5. BUSA, Roberto. Foreword: Perspectives on the Digital Humanities. In *A Companion to Digital Humanities*. Ed. by Susan Schreibman; Ray Siemens; John Unsworth. Kindle Edition. Oxford: Blackwell, 2004. Kindle Locations 187–188.
6. DIEM, Markus, Robert Sablatnig, Melanie Gau, Heinz Miklas. Recognizing Degraded Handwritten Characters. In *Kodikologie und Paläographie im digitalen Zeitalter 2 = Codicology and Palaeography in the Digital Age 2.* Ed. by Franz Fischer, Christiane Fritze, Georg Vogeler. Norderstedt: BoD, 2010, p. 295–306.
7. DUNSIRE, Gordon. Linked data for manuscripts in the Semantic Web. In *Summer School in the Study of Historical Manuscripts*: Proceedings. Ed. by Mirna Willer and Marijana Tomić. Zadar: Sveučilište, 2013, p. 199–218 [accessed 4 August 2013]. Access (to pre-print) through Internet: <http://www.gordondunsire.com/pubs/docs/LinkedDataForManuscripts.pdf>.
8. HERCIGONJA, Eduard. *Tropismena i trojezična kultura hrvatskoga srednjovjekovlja*. Zagreb: Matica hrvatska, 1994.
9. JAKOBI-MIRWALD, Christine. *Buchmalerei*: ihre Terminologie in der Kunstgeschichte. Berlin: Reimer, 1997.
10. *Lexicon der germanistischen Linguistik*. Ed. Alhaus, H. P.; H. Henne, H. E. Wiegand. Tübingen, 1973.
11. MANIACI, Marilena. *Terminologia del libro manuscritto*. Roma: Istituto per la patologia del libro; Editrice Bibliografica, 1997.
12. MUZERELE, Denis. *Vocabulaire codi-cologique*: répertoire méthodique des termes français relatifs aux manuscrits. Paris: Editions CEMI, 1985 [accessed 4 August 2013]. Access through Internet: <http://vocabulaire.irht.cnrs.fr/pages/vocab2.htm>.
13. *Open Metadata Registry* [accessed 4 August 2013]. Access through Internet: http://metadataregistry.org/.

14. PIERAZZO, Elena; Peter Stokes. Putting the text back into Context: a Codicological Approach to Manuscript Transcription. In *Kodikologie und Paläographie im digitalen Zeitalter 2 = Codicology and Palaeography in the Digital Age 2*. Ed. by Franz Fischer, Christiane Fritze, Georg Vogeler. Norderstedt: BoD, 2010, p. 397–430.
15. PELC, Milan. *Pismo – knjiga – slika*: uvod u povijest informacijske kulture. Zagreb: Golden marketing, 2002.
16. RENHART, Erich. Manuscript fragments: the hidden library. In *Summer school in the Study of Historical Manuscripts, Zadar, Croatia, 26–30 September 2011*. Ed. by Mirna Willer and Marijana Tomić. Zadar: Sveučilište, 2013.
17. *Resource description Framework*: RDF [accessed 04 August 2013]. Access through Internet: <http://www.w3.org/RDF/>.
18. STOKES, Peter. Palaeography and image processing: some solutions and problems. *Digital Medievalist*, 2007/2008, no. 3 [accessed 04 August 2013]. Access through Internet: <www.digitalmedievalist.org/journal/3/stokes/>.
19. SVENSSON, Patrik. Humanities Computing as Digital Humanities. *Digital Humanities Quarterly*, 2009, 3, no. 3 [accessed 4 August 2013]. Access through Internet: <http://digitalhumanities.org/dhq/vol/3/3/000065/000065.html>.
20. SVENSSON, Patrik. The landscape of Digital Humanities. *Digital Humanities Quarterly* 2010, 4, no. 1 [accessed 04 August 2013]. Access through Internet: <http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html>.
21. VLAŠIĆ-ANIĆ, Anica. Novootkriveni glagoljski fragmenti u riječkoj kapucinskoj knjižnici. In *Sprache und Leben der frühmittelalterlichen Slaven*. Ed. Stadnik-Holzer, Elena; Georg Holzer. Frankfurt am Main Peter Lang, 2010, p. 95–124.
22. VLAŠIĆ-ANIĆ, Anica. Fragments of parchment codices on the covers of old books. In *Summer school in the Study of Historical Manuscripts*, Zadar, Croatia, 26–30 September 2011. Ed. by Mirna Willer and Marijana Tomić. Zadar: Sveučilište, 2013, p. 145–172.
23. WILLER, Mirna; DUNSIRE, Gordon. *Bibliographic Information Oganization in the Semantic Web*. Cambridge: Chandos, 2013 (October).
24. W3C. SKOS Simple Knowledge Organization System Primer, W3C Working Group Note 18, August 2009 [accessed 04 August 2013]. Access through Internet: <www.w3.org/TR/skos-primer>.
25. ŽAGAR, Mateo. *Grafolingvistika srednjovjekovnih tekstova*. Zagreb: Matica hrvatska, 2007.

SKAITMENINĖS HUMANITARIKOS IŠPLĖTIMO SEMANTINIAME ŽINIATINKLYJE GALIMYBĖS: KROATIJOS VIDURAMŽIŲ RANKRAŠČIŲ, INKUNABULŲ IR JŲ FRAGMENTŲ ATVEJO ANALIZĖ

Marijana Tomić, Mirna Willer

Santrauka

Rankraščių rinkiniai – tai labai įvairaus pobūdžio rankraščiai, paprastai apibrėžiami kaip „ranka ant popieriaus arba pergamento užrašytas tekstas arba dokumentas“ (Peter Beal). Tai gali būti šeimos ar asmeniniai dokumentai, dienoraščiai, laiškai, archyvų rinkiniai ir kt. Viduramžių rankraščiai – kodeksai, žemėlapiai, muzikos kūriniai arba jų fragmentai – sudaro specialią rankraščių rūšį. Kaip ir inkunabulai, rankraščių rinkiniai yra vertingiausia bibliotekų paveldo dalis, dėl jų mus pasiekia itin daug informacijos apie viduramžių istoriją, kultūrą, literatūrą, socialinę istoriją,

gyvenimo tendencijas. Be šių šaltinių informacija būtų dingusi. Senų ir retų rankraščių tyrimai svarbūs tiek šalies, tiek visos Europos kultūros ir socialinei istorijai.

Žvelgiant iš humanitarinių mokslų perspektyvos, būtina išskirti keletą veiksnių, kurie lėmė reikšmingus pokyčius tyrinėjant rankraščius ir pirmąsias spausdintines knygas. Pačiu svarbiausiu laikomas informacinių technologijų poveikis beveik visoms tyrimo sritims. Šie pokyčiai lėmė ir naujos disciplinos – skaitmeninių humanitarinių mokslų atsiradimą. Pasak Toby'o Burrowso, viduramžių tyrinėtojai yra „pažangiausi skaitmeninių technologijų taikymo humanitarinių mokslų tyrimuose atstovai". Vis dėlto T. Burrowsas išskiria ir keletą keblumų, susijusių su interneto ir skaitmeninės bibliotekos paslaugomis. Jis nurodo „integracijos ir sąveikos tarp daugybės skirtingų interneto svetainių stygių" bei terminologijos nenuoseklumą taikant aprašomuosius standartus. Savo ruožtu tai sukelia probleminę situaciją, nes „tyrinėtojams visame pasaulyje kyla daug sunkumų rasti, naudotis ir dalytis žiniomis apie viduramžių rankraščių kolekcijas". Visiškai pritariame T. Burrowso minčiai, kad šią problemą galima išspręsti sukuriant tarptautinę bendradarbiavimo infrastruktūrą, kuri leistų tvarkyti turinį ir tarpusavyje susijusias žinias. Mūsų nuomone, ši infrastuktūra gali būti įgyvendinta technologinėje semantinio žiniatinklio ir sujungtų atvirų duomenų (angl. *Semantic Web and Linked Open Data*) terpėje.

Straipsnyje aptariami viduramžių rankraščių ir inkunabulų bei jų fragmentų tyrimai ir šių šaltinių aprašymas kaip skaitmeninių humanitarinių mokslų projekto dalis, taikant šią naują technologiją. Nagrinėjamas šios srities Kroatijos Zadaro universiteto Informacijos mokslų fakulteto vykdomas mokslinių tyrimų projektas. Projekto tikslas – atrinkti duomenų elementus, reikalingus tiksliam minėtų šaltinių aprašymui ir jų standartizavimui, naudojant senų ir retų knygų tyrinėtojų parengtas bibliografijos, kodikologijos, paleografijos bei tipografijos ontologijas.

Straipsnyje pateikiamas ir trumpas technologinės semantinio tinklo infrastruktūros bei jo standartų įvadas. Detaliai aprašoma metodika, padedanti paskelbti pasirinktą žodyną kaip vieną iš metaduomenų registro paslaugų. Pateikiamas sujungtų atvirų duomenų paskelbimo pavyzdys – pristatatomas grafikas, vaizduojantis iš dalies rekonstruoto rankraščio fragmento aprašymą. Kadangi visos minėtos disciplinos naudoja savo žodynus ir ontologijas, straipsnio autorės siūlo orientuotis ne į vieno bendro žodyno naudojimą, o į atitinkamų terminų sąsajų projektavimą vadovaujantis SKOS taisyklėmis. Taip būtų kuriami būsimos tarptautinės bendradarbiavimo struktūros pagrindai.

Įteikta 2013 m. rugpjūčio mėn.