Dublin Core Metadata for Research Data – Lessons Learned in a Real-World Scenario with datorium

Andias Wira Alam

GESIS – Leibniz Institute for the Social Sciences

Germany

andias.wira-alam@gesis.org

Abstract

As a continuation of our work in the datorium project, we provide a service for autonomous documentation and upload of research data. In this paper we discuss and share our experience of developing such a service by using Dublin Core Metadata. Even small and simple, DC Metadata is an appropriate standard to be taken as basic metadata, for instance in the repository systems. The required elements for describing research data are mostly complex, in particular the acquired information about the data, including survey methods, survey periods, or number of variables. DC Metadata cannot cover all elements needed in the research data repository. However, we show that with some extended elements and front-end based manipulations the DC Metadata can be applied usefully in this real-world scenario and support complex description without overcoming the "simplicity" of the standard.

Keywords: research data repository; metadata; DSpace; infrastructure; datorium

1. Introduction

GESIS – Leibniz Institute for the Social Sciences provides services for research in multiple phases of the research process, such as study planning, data collection, data analysis, and archiving and registration. The main targets are data collected from surveys, data from historical social research, as well as scientific publications. Figure 1 shows the research data lifecycle used by the institute to structure its services. Our project datorium belongs to the phase "archiving and registering". We provide a data repository service for social science and economic researchers with a user-friendly tool for the autonomous documentation, upload and publication of their research data, as illustrated in Figure 2. As stated in Linne (2012), the service focuses particularly on small research projects by researchers who do not necessarily work for an institution or are self-funded. A detailed review carried out by the GESIS Data Archive ensures the quality of the submitted data.

Describing research data requires comprehensive and rich metadata elements, such as provided by the Data Documentation Initiative (DDI)[1] or da|ra metadata[2]. DDI can be used not only to describe the research data on study level (general overview of the research data), but also on the variable level - e.g. for information details about variables, questionnaires, and results. The da|ra metadata is now commonly used in assigning persistent identifiers to research data in the context of the DataCite[3] community. Nonetheless, DC elements are the most-used elements for describing resources, particularly scientific resources (cf. Ell et al. 2011; Qin et al. 2013; Malta et al. 2014). Príncipe et al. (2013) also stated that OpenAIRE is starting to move from a publication infrastructure to a more comprehensive infrastructure that covers all types of scientific "output". DC metadata as a fundamental metadata infrastructure for scientific publications is therefore slowly evolving into an infrastructure for research data as well.

---

[1] See http://www.ddialliance.org/
[2] See http://www.da-ra.de/en/home/
[3] http://www.datacite.org/

◆DCPAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2014

The choice of metadata schema or standard is likely not the main focus for researchers looking to publish the data. Researchers need platforms which allow them to publish their data in an easy way, making the data visible and citable (Wira-Alam et al. 2012; Dimitrov et al. 2013). During the requirements analysis for datorium it became apparent that we had to specify the requirements so as to balance simplicity and usefulness. Similar to the lessons learned reported by Wallis et al. (2010), though in a different context, the discussions between the computer and social scientists in the project team started with the question "what should we build for you?" answered by "what could you build for us?".
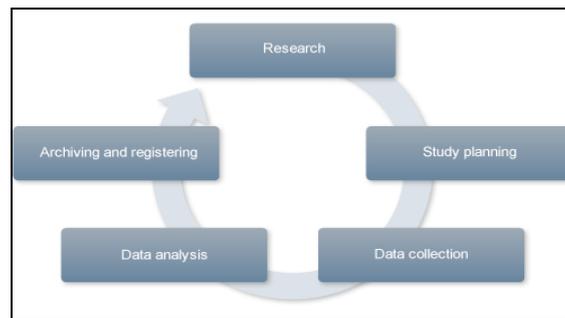

FIG. 1. Research data lifecycle in multiple phases[4]
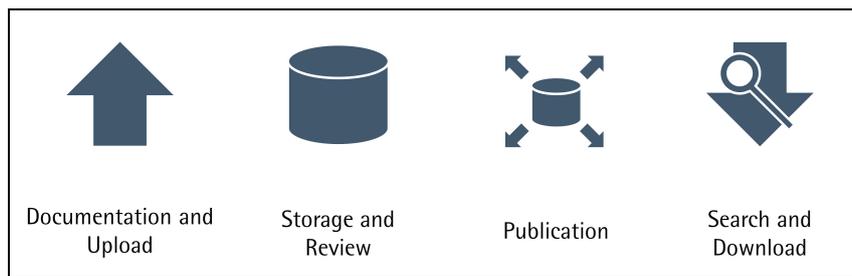
FIG. 2. Illustration of the step-by-step processes within datorium[5]

An ideal vision is that any piece of information in the research data should be well-documented and described. Castro et al. (2013) proposed e.g. to use domain-specific elements in order to fully describe scientific experiments. However, our tool is targeted not only at institution-based researchers but also at any self-funded researchers or even students. Documenting and describing research data is time-consuming and hence expensive work. Thus, increasing the complexity of the documentation process would impact the usability of the tool, and consequently potential users might lose their interest in using it. Accordingly, one of the key challenges is how to make the tool as simple as possible for users, in particular the data depositors, while at the same time gathering as much information about the data as possible. Simultaneously, we have to make the data visible and easy find, especially for the data consumers. Another important feature of the tool is that it shall be available in two languages, namely German and English, in order to target prospective international users: data depositors as well as data consumers.

2. Metadata Design

In the metadata design, we identify not only critical information about metadata in general, but also more detailed information about the research data, e.g. survey / data collection methods, survey periods, or number of variables / units. However, in order to keep the metadata simple, we

---

[4] http://www.gesis.org/en/services/
[5] https://datorium.gesis.org/

DCPAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2014

use a Dublin Core subset whose schema is simply flat and has no complex hierarchical structure. As stated by Rice et al. (2008) and Greenberg et al. (2013), datasets are digital materials that need to be described for discovery, preservation, and re-use, e.g. for partner repositories. Furthermore, analogous to the aforementioned work and Greenberg et al. (2009), by using DC elements we provide understandable information about complex objects and help partner repositories or data consumers to become acquainted with the research data. Research data also become more useful when they are interoperable with other data and therefore need a common standard or set of standards (Ball, 2010). As depicted in Figure 3, datorium's metadata schema consists of DC elements and some extended elements. As some elements, e.g. file description, demand hierarchical entries, the schema forms a tree structure. A complex tree structure cannot be described by a schema such as Dublin Core. As mentioned in Chen et al. (2013), research datasets may contain unusual file formats; therefore the uploaded files need additional information e.g. on the number of variables, number of units, languages used in the files, or even software to read the files for further processing. Figure 3 depicts the abstraction of the metadata schema.
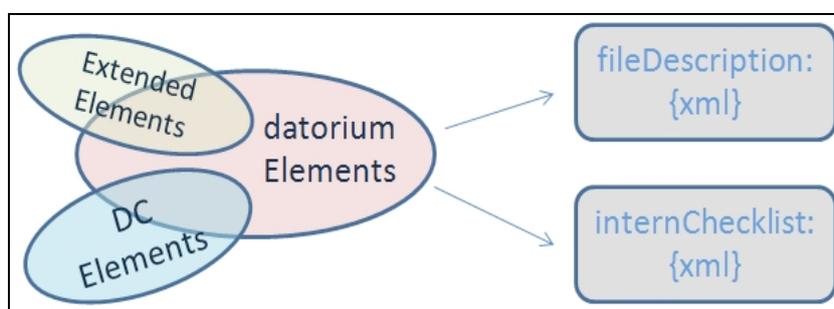

FIG. 3. Illustration of the metadata schema in datorium

To meet the requirement that we need as much information about the data as possible We require 6 mandatory and about 14 optional entries. As mentioned above, we assume that most users are not willing to capture many entries in the tool for simplicity reasons. However we cannot exclude this possibility as there are users who want to provide rich information about their data e.g. to increase the visibility of the data. This situation contrasts with the identified requirements, but we discuss later in the next section how we alleviate this problem. In Table 1 we describe our metadata schema. In comparison with the first design (Wira-Alam et al, 2012), we use 10 DC elements and specify all extended elements with the namespace "dbk" taken from GESIS – Data Catalogue DBK[6]. We also organized the elements in five groups, e.g. General Description or Methodology, according to their affinities. Moreover, we decomposed two elements, Primary Researcher and Contributor, to increase the exactness. In the element Primary Researcher, for instance, we distinguish between person and institution. According to DC standard, however, this property can be filled either with person or institution. Our adaptation makes it possible for users to search only for persons or institutions.

As mentioned above, we support users by increasing the visibility of the data. For this purpose we offer controlled vocabularies, e.g. for Subject Area or Data Collection Method. Controlled vocabularies improve the visibility of the data on the one hand by enhancing the semantic of the metadata, and on the other hand by making the submission process easier for the users. Moreover, in order to support internationalization, we provide all controlled vocabularies in two languages: German and English. This affects both the technical implementation and the search functionality. We demonstrate in the next section how users can benefit from this feature and what the technical implementation looks like.

---

[6] https://dbk.gesis.org/

DC PAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2014

TABLE 1: Metadata schema using DC elements and extended elements

	Labels	DC Elements	Extended Elements	Type
General Description	*Title*	dc.title	-	*text*
	Other Title *	-	dbk.othertitle 　dbk.othertitle.text 　dbk.othertitle.type　　　　*(a)*	 *text* *text*
	DOI	dc.identifier.uri	-	*URI*
	*Primary Researcher **	dc.creator	dbk.primaryresearcher.person dbk.primaryresearcher.institution	*text* *text*
	Publisher	dc.publisher	-	*text*
	Publication Year	-	dbk.publicationyear	*date: YYYY*
	Availability Embargo 　Embargo (until)	- - -	dbk.availability　　　　　　*(b)* dbk.embargo.availability dbk.embargo.end	*text* *text* *date: YYYYMMDD*
	Contributor *	dc.contributor	dbk.contributor.person dbk.contributor.institution dbk.contributor.type　　　　*(c)*	*text* *text* *text*
Content	Subject Area *	dc.subject.other	-　　　　　　　　　　　　*(d)*	*text*
	Topic Classification *	dc.subject.classification	-　　　　　　　　　　　　*(e)*	*text*
	Abstract	dc.description	-	*text*
Methodology	Geographical Area *	dc.coverage.spatial	-　　　　　　　　　　　　*(f)*	*text*
	Universe *	-	dbk.universe	*text*
	Selection Method	-	dbk.selectionmethod	*text*
	Data Collection Method *	-	dbk.datacollectionmethod　*(g)*	*text*
	Survey Period *	-	dbk.surveyperiod 　dbk.surveyperiod.start 　dbk.surveyperiod.end	 *date* *date*
Additional Notes	Rights *	dc.rights	-	*text*
	Notes *	-	dbk.notes 　dbk.notes.text 　dbk.notes.type　　　　　*(h)*	 *text* *text*
	Source *	-	dbk.source	*text*
	Publications *	-	dbk.publication 　dbk.publication.text 　dbk.publication.id	 *text* *text*
Files	File *	-	dbk.file 　dbk.file.filename 　dbk.file.filedescription 　dbk.file.version 　　dbk.file.versionNumber 　　dbk.file.versionDate 　dbk.file.resource 　　dbk.file.resourceType 　　dbk.file.resourceTypeGeneral 　dbk.file.language　　　　*(i)* 　dbk.file.numberofvariables 　dbk.file.unit 　　dbk.file.unitNumberOf 　　dbk.file.unitType　　　*(j)* 　dbk.file.software 　dbk.file.alternateId 　dbk.file.relatedId 　　dbk.file.relatedIdIdentifier 　　dbk.file.relatedIdType　*(k)*	 *text* *text* *text* *text* *date: YYYYMMDD* *text* *text* *text* *text* *int* *text* *int* *text* *text* *text* *text* *text* *text*
Hidden	*Date Issued (for sorting purpose)*	dc.issued	-　　　　　　　　　　　　*(l)*	*date: YYYYMMDD*
	Checklist (for Curators only)	-	intern.cheklist	*text*

As explained, the elements consist of DC elements and DBK elements. In the first group, namely General Description, we place all mandatory entries (written in italics). On publication each submission automatically receives a persistent identifier, in this case a DOI® (generated by the system). This increases the visibility of the submitted data by making them as they are citable as scientific publications. We set GESIS – Data Archive as the publisher of the data since they are published through datorium; therefore the value of the element has been fixed and it is not editable. Elements marked with an asterisk are repeatable. We also introduce Other Title in order to accommodate research data that have several titles for some reasons, e.g. original title, translated title in several languages, or project title.

In the second group, namely Content, users can provide a description of the data as a free-text abstract. In addition, users have two important elements: Subject Area and Topic Classification, whose values are controlled vocabularies. The controlled vocabularies provide a possibility for semantic enhancement and thus facilitate connections between the research data and Linked Data on the Web (cf. Isaac et al. 2013). In the group Files we collect relevant information about the files as completely as possible. Further explanation for the elements marked alphabetically from *(a)* to *(l)* is as follows:

(a) dbk.othertitle.type – Type of other title can be selected from the controlled vocabularies provided by DBK (Zenk-Möltgen et al. 2012), such as "project title" or "original title".

(b) dbk.availability – It consists of three controlled vocabularies: "free access", "restricted access", and "embargo".

(c) dbk.contributor.type – Type of contributor is based on the category scheme of the ContributorType from DataCite.

(d) dc.subject.other – Subject Area has been chosen from the disciplines in SSOAR - Social Science Open Access Repository.

(e) dc.subject.classification – Topic Classification is based on DBK[7], consists of overall 38 terms, such as "Economic Systems" or "Social Policy".

(f) dc.coverage.spatial – Geographical Area consists of places, such as countries, cities, or provinces / states, based on ISO-3166.

(g) dbk.datacollectionmethod – Data Collection Method consists of the controlled vocabularies provided by DDI (unreleased beta version, March 2013), such as "Email interview", "recording", or "Telephone interview: CATI".

(h) dbk.notes.type – It is based on DescriptionType provided by DataCite, such as "Abstract" or "TableOfContents".

(i) dbk.file.language – Languages provided by ISO-639.

(j) dbk.file.unitType – Unit Type is based on "Analysis Unit" provided by DDI, such as "Family", "Individual", or "Organization".

(k) dbk.file.relatedIdType – Type of the related identifier is based on the RelationType provided by DataCite, such as "IsCitedBy", "IsDocumentedBy", or "IsPartOf".

(l) dc.issued – Date Issued has been generated by the system at the time of publication.

As mentioned above, datorium offers multi-language support for the controlled vocabularies. The tool supports a so-called ad-hoc translation automatically. Users do not have to take any action in this regard. All controlled vocabularies are stored in a dictionary in two languages. Each vocabulary item in both languages is unique and therefore the correctness of the translation is guaranteed. The controlled vocabularies for Subject Area and Data Collection Method have a tree structure, in opposite of having a long list, to make it easier for users to find and choose the relevant terms for their data.

For the types of the elements, we use rudimentary types for reasons of simplicity. Thus, there are only 4 rudimentary types: *text*, *URI*, *int*, and *date*. Theoretically, with *text* we can cover any types of values. However, we apply a simple validation in order to avoid wrong values. Values typed with *date* without a fixed date format, namely Survey Period, can be given in three

---

[7] https://dbk.gesis.org/dbksearch/Kategorien.htm

variants: year only (format: YYYY), month and year only (format: YYYYMM), or an exact date (format: YYYMMMDD).

3. Technical Implementation

We use a DSpace[8] repository as basis platform for the implementation. The metadata model in DSpace, which is based on Dublin Core, is simply flat and has no complex hierarchical structure. It consists of *schema*, *element*, and *qualifier*. A *schema* is equivalent to namespace, *element* can be considered as content, and *qualifier* can be seen as sub-element if an extra attribute needs to be added. DSpace is a web-based application that follows the Model-view-controller (MVC) architectural pattern (Gamma et al. 1994). This pattern ensures the consistency of the model (data) and the user interface / front-end (view) by employing a controller. DSpace also offers many features such as user management, review process, and discovery / faceted search. Our development process is loosely based on agile software development, which is an iterative process throughout the development cycle.

As described in Table 1, we have groups of elements. In the implementation, we display each group of elements as a tab. This strategy is suitable for data depositors who do not want to spend time capturing information about the data. However, even though all mandatory elements are placed in the first tab, each submission needs to go through all. Figure 4 shows the mandatory and non-mandatory elements in the first tab. For example, the mandatory element Principle Investigator can be filled only by a person, an institution, or both. For the data depositors who are willing to provide as more information about their data, this strategy is also convenient as it provides more structure and orientation for data depositors than a single form with many elements. After the data has been successfully published, the system will assign a persistent identifier (DOI) automatically via a separate module connected with the da|ra API for DOI registration[9].

FIG. 4. Editor form for General Description

---

[8] As we mentioned in the previous work (Wira-Alam et al. 2012), we use DSpace (version 1.8.2) as it is an open source repository application. Furthermore, DSpace supports Dublin Core elements by default and has a flat metadata schema which helps us as developers to maintain the data. According to DSpace's website, there are more than 1000+ institutions that have registered to use DSpace for their repository application which is widely used worldwide (May 2014).

[9] http://www.da-ra.de/en/for-data-centers/register-data/

In the Content tab, as captured in Figure 5, one important feature of the tool, namely the controlled vocabularies is shown. We display the vocabularies in their original, i.e. hierarchical form. The hierarchical selection is very comfortable since users can, for example, find or determine an appropriate subject, or more, by its discipline. This feature was implemented without changing the metadata schema. We performed a pure front-end based manipulation and thus the validation occurs in the view as well. A big advantage of this strategy is that the metadata schema becomes flexible since the view does not depend on the model. A possible disadvantage could be wrong values in the database because of a front-based validation level that does not guarantee the consistency. Nevertheless, wrong values only apply for the corresponding element and cannot break the whole elements. Besides, a review process is carried out before the data is published.
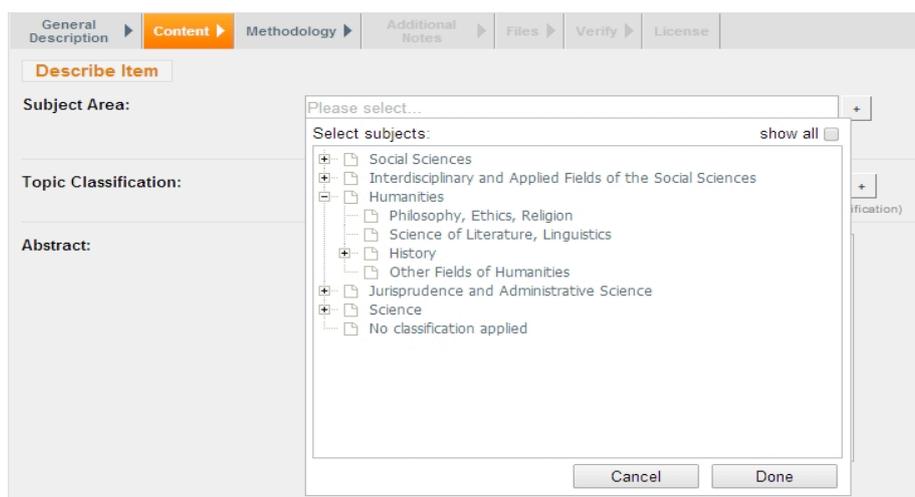
FIG. 5. Editor form for Content

The next feature regarding the controlled vocabularies is autocomplete. In Figure 6, in the element Geographical Area the data depositors can select places from a given list. Since there are thousands of places to be selected, we provide an autocomplete widget in order to make the selection easier. Users can decide the preferred language (DE/EN); the whole user interface and the controlled vocabularies are then available in the selected language. Another feature is a widget to pick a date. This can be an exact date but also year only or month and year only.
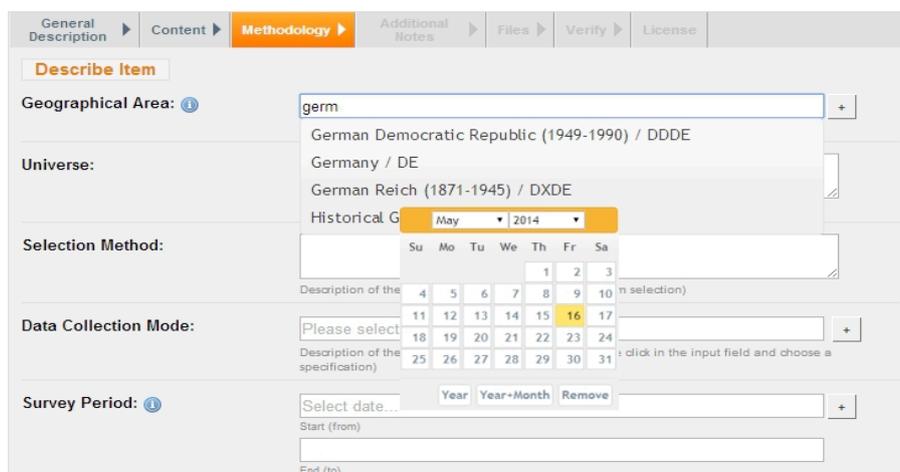
FIG. 6. Editor form for Methodology

DCPAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2014

As repeatedly mentioned, since we cannot apply a complex metadata schema in the model, we can only modify the view to meet the requirements. For instance, each uploaded file has several elements and each submission / dataset can have several files since it is a repeatable element. This situation therefore leads to a hierarchical form in the model, which is actually not implementable. As shown in Figure 7, we *wrap* these elements in an XML as if they are seen as a single value of the element File to compromise the limitation of the flat metadata model.

FIG. 7. Editor form for Files Upload

For the data consumers, finding data is an intellectual effort. In addition to free-text search, faceted search is a well-known technique that helps users to browse large data collections, e.g. images or documents, and delve into more details if required (Yee at al. 2003). By using this technique, and since the controlled vocabularies are available in German and English, the data can be also searched with keywords in a language in which the data was not documented originally. Figure 8 demonstrates the multi-language support for the faceted search. The element Geographical Area shows same values according to the preferred language.

FIG. 8. Multi-language support for faceted search

●**DC**PAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2014

All front-end based manipulations make use of JavaScript, in particular jQuery[10] and its plugins, and had been successfully tested in various browsers in different versions, among others Internet Explorer, Safari, Opera, Mozilla Firefox, and Google Chrome. The layout and user interface are based on Manakin's XMLUI[11] with many modifications according to the GESIS Web-Style-Guide[12].

4. Conclusion and Future Work

We use Dublin Core for our purpose since it is a simple and appropriate schema for documenting research data. However, to meet all requirements, some extensions are needed. We have shown some approaches to make the application useful and cover complex description without overcoming the "simplicity" of DC metadata. The front-end based manipulation, as we demonstrated in this paper, can remedy the limitation of the schema, e.g. to deal with complex, repeatable elements structures. The documentation of the research data currently refers to the study level; details about the variables used in the survey are not covered. Nevertheless, it is at all times possible to extend the schema so as to meet new requirements. Since the schema and front-end are quite distinct from each other, our approach is suitable for this situation because of its flexibility.

As future work, we want to establish the connection between publication and research data automatically (Boland et al. 2012; Ritze et al. 2013) in order to incorporate scientific publications in research data and the other way around. Moreover, an integrated search with other partner repositories is under way. Therefore we plan to implement an export / import, harvesting (e.g. OAI-PMH) interface, and a schema crosswalk to other standards, e.g. DDI.

Acknowledgements

References

Ball, A. (2010): Review of the State of the Art of the Digital Curation of Research Data (version 1.1). ERIM Project Document erim1rep091103ab11. Bath, UK: University of Bath. Retrieved April 28, 2014 from http://opus.bath.ac.uk/18774/2/erim1rep091103ab11.pdf

Boland, K., Ritze, D., Eckert, K., Mathiak, B. (2012): Identifying References to Datasets in Publications. TPDL, Vol. 7489 of Lecture Notes in Computer Science, page 150-161. http://dx.doi.org/ 10.1007/978-3-642-33290-6_17

Castro, J. A., Ribeiro, C., Rocha da Silva, J. (2013): Designing an Application Profile Using Qualified Dublin Core: A Case Study with Fracture Mechanics Datasets. Proc. International Conference on Dublin Core and Metadata Applications 2013. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/view/3685

Chen, H., Lin, Y., Chen, C. (2013): Approaches to Building Metadata for Data Curation. Proc. International Conference on Dublin Core and Metadata Applications 2013. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/view/3691

Dimitrov, D., Baran, E., Wegener, D. (2013): Making Data Citable - A Web-based System for the Registration of Social and Economics Science Data. In: Krempels, Karl-Heinz; Stocker, Alexander (Hrsg.): Proceedings of the 9th

---

[10] http://jquery.com/
[11] https://wiki.duraspace.org/display/DSDOC18/XMLUI+Configuration+and+Customization
[12] http://www.gesis.org/styleguide/

DCPAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2014

International Conference on Web Information Systems and Technologies: Aachen, Germany, 8 - 10 May 2013: SciTePress, pages 155-159

Ell, B., Vrandečić, D., Simperl, E. (2011): Labels in the Web of Data. In Proceeding of the 10th International Semantic Web Conference, 2011. http://dx.doi.org/10.1007/978-3-642-25073-6_11

Gamma, E., Helm, R., Johnson, R., Vlissides, J. (1994): Design Patterns: Elements of Reusable Object-Oriented Software. Addison-Wesley Professional. ISBN-13: 978-0201633610

Greenberg, J., Swauger, S., Feinstein, E. (2013): Metadata Capital in a Data Repository. Proc. International Conference on Dublin Core and Metadata Applications 2013. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/view/3678

Greenberg, J., White, H. C., Carriera, S., Scherleb, R. (2009): A Metadata Best Practice for a Scientific Data Repository. Journal of Library Metadata. Volume 9, Issue 3-4, pages 194-212. doi:10.1080/19386380903405090

Isaac, A., Charles, V., Fernie, K., Dallas, C., Gavrilis, D., Angelis, S. (2013): Achieving Interoperability between the CARARE Schema for Monuments and Sites and the Europeana Data Model. Proc. International Conference on Dublin Core and Metadata Applications 2013. http://dcevents.dublincore.org/IntConf/dc-2013/paper/view/171

Linne, M. (2013): Sustainable data preservation using datorium: facilitating the scientific ideal of data sharing in the social sciences. In: Borbinha, José; Nelson, Michael; Knight, Steve (Hrsg.): Proceedings of the 10th International Conference on Preservation of Digital Objects, Lisbon: Biblioteca Nacional de Portugal, S. 150-155. Retrieved May 16, 2014 from http://purl.pt/24107/1/

Malta, M. C., Baptista, A. A. (2014): A panoramic view on metadata application profiles of the last decade. Int. Journal of Metadata Semantic and Ontologies Vol. 9, Issue 1 (February 2014), pages 58-73.

Príncipe, P., Rodrigues, E., Rettberg, N., Schirrwagen, J., Loesch, M., Karstensen, M., Nielsen, L. H. (2013): OpenAIRE Guidelines for Data Archive, Literature Repository and CRIS Managers. Proc. International Conference on Dublin Core and Metadata Applications 2013. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/viewFile/3695

Qin, J., Li, K. (2013): How Portable Are the Metadata Standards for Scientific Data? A Proposal for a Metadata Infrastructure. Proc. International Conference on Dublin Core and Metadata Applications 2013. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/viewFile/3670/1893

Rice, R. (2008). Applying DC to Institutional Data Repositories. Proceedings of the International Conference on Dublin Core and Metadata Applications, 2008. Retrieved May 16, 2014 from http://dcpapers.dublincore.org/pubs/article/view/945

Ritze, D., Boland, K. (2013): Integration of Research Data and Research Data Links into Library Catalogues. Proc. International Conference on Dublin Core and Metadata Applications 2013. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/view/3683

Wallis, J. C., Mayernik, M. S., Borgman, C. L., Pepe, A. (2010): Digital libraries for scientific data discovery and reuse: from vision to practical reality. Proc. 10th Joint Conference on Digital libraries (JCDL '10). ACM, New York, NY, USA, pages 333-340. doi:10.1145/1816123.1816173

Wira-Alam, A., Dimitrov, D., Zenk-Möltgen, W. (2012): Extending Basic DublinCore Elements for an Open Research Data Archive. Proc. International Conference on Dublin Core and Metadata Applications 2012. Retrieved April 28, 2014 from http://dcpapers.dublincore.org/pubs/article/view/3664/1887

Yee, K.-P., Swearingen, K., Li, K., Hearst, M. (2003): Faceted Metadata for Image Search and Browsing. ACM SIGCHI: Human Factors in Computing Systems, 2003. http://dx.doi.org/10.1145/642611.642681

Zenk-Möltgen, W., Habbel, N. (2012): Der GESIS Datenbestandskatalog und sein Metadatenschema. Version 1.8. GESIS Technical Reports 2012/1. Retrieved June 21, 2012 from http://nbn-resolving.de/urn:nbn:de:0168-ssoar-292372