Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

13th International Conference on Current Research Information Systems, CRIS2016, 9-11 June 2016, Scotland, UK

Modelling National Research Information Contexts based on CERIF

Christoph Quix[a,b], Mathias Riechert[c,d]

[a] *Fraunhofer Institute for Applied Information Technology FIT, 53754 St. Augustin, Germany*
[b] *RWTH Aachen University, 52056 Aachen, Germany*
[c] *German Centre for Higher Education Research and Science Studies (DZHW), 10117 Berlin, Germany*
[d] *Technical University of Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany*

Abstract

We present the results of the data modelling in the project 'Research Core Dataset' (2013-2015) that provided the context for developing a set of core definitions for research information for the German science system. In this paper, we focus on the data modelling aspects of the project, whereas another submission focuses on the management of the discussion phase and visualization of the argumentation process in the project. We present how the data model has been developed and synchronised with the argumentation process. As compatibility with CERIF was a major requirement for the data model, we present our approach to link the data model to the CERIF standard.

Peer-review under responsibility of the Organizing Committee of CRIS2016.

Keywords: research information; CERIF; data modelling

1. Purpose

Modelling a standardised data model for a national context of research information is a complex challenge, especially if the interests of hundreds of stakeholders in research information management are to be taken into account. Within the German project for defining a 'Research Core Dataset' for the German science system this national context has been specified for the German Science System[1]. In this paper, we report the results with respect to the data modelling part and its relationship to CERIF. A related paper presents the framework for documenting the standardisation process[2].

Harmonising the reporting of research information on a national level requires the agreement of the participating stakeholders and 'clearly defined requirements' for a sustainable conceptual model[3]. To achieve this goal, we characterised the standardisation process for research information as a wicked problem[4,5], i.e., a problem that does not have a clear definition and its solution cannot be true-or-false, but only good-or-bad. For wicked problems, the focus is rather on the problem-solving process instead of the final solution[6]. The solution of wicked problems involves various stakeholders who exchange arguments pro and con certain issues.

E-mail address: christoph.quix@fit.fraunhofer.de

Peer-review under responsibility of the Organizing Committee of CRIS2016.

Thus, the visualisation of the argumentation and discussion process is an important issue for the acceptance of a data model for research information. In[7], we presented a framework in which techniques from CSAV (Computer Supported Argumentation Visualization,[8]) and IBIS (Issue-based Information Systems,[9]) are combined to produce a comprehensive but still comprehensible documentation of the argumentation process.

The result of the framework is a set of definitions for elements of a data model for research information in a semi-structured form[a]. The arguments are linked to the corresponding definitions. Although the elements and their definitions have been captured in a tabular form, the definitions are still insufficient for a formal exchange of research information. Therefore, the project included a working group that aimed at the translation of the definitions into a formal data model, which should guarantee at the same time that the new standard can be implemented with reasonable efforts in current research information systems.

2. Approach

From the perspective of a data modeller, the definitions are the requirements for the data model and, thus, form an important input for the data modelling process. The data modelling process was done in a classical way: we created first a conceptual data model and then mapped it to a logical data model[10]. But why is it necessary to create new data models for research information although there is already the CERIF standard?

2.1. Limitations of existing standards

CERIF provides a wide-ranging model for research information. It provides definitions for most elements which are relevant for research information management. However, as a European standard, CERIF cannot provide all the details which might be relevant only in one country. This requirements-driven approach has also been pointed out in a recent study that emphasised the need to standardise research contexts in contrast to only standardise research information[3].

Furthermore, the definitions in CERIF are rather ‘abstract’ as the standard should be applicable to several use cases. A specific contextual semantics that defines the elements with respect to a given application context is not part of the standard[3]. For example, it has been not specified which kind of projects should be considered as an instance of ‘cfProject’ or which prizes are relevant for ‘cfPrize’.

A similar question as for CERIF could be also raised for CASRAI. CASRAI maintains and develops profiles for research administration information[3]. The process of defining research information based on business needs is similar to our approach to model research information based on requirements. At the time of our project, the existing definitions in CASRAI and their underlying business needs did not match the requirements for the German Science System which expected more detailed definitions and data models.

The definition of these contextual boundaries was the main goal of the discussion and argumentation phase. This included, for example, the discussion about which staff categories should represented in the German core dataset, or which types of projects should be considered. In addition, classification, categorisation, and attributes of the research information entities have been discussed.

2.2. Representation of the Definitions in the German Research Core Dataset

As mentioned above, this information has been captured in a semi-structured form, which could be visualised in tables[b], argumentation graphs, and other more aggregated views[7,2].

These different views were very useful for the data modelling as with a more abstract view, the main areas and their relationships could be identified; on the other hand, modelling details of research information required also the

[a] The resulting table is available at `http://kerndatensatz-forschung.de/version1/Spezifikationstabelle_KDSF_v1.html`.

[b] In addition to the table with the final definitions (`http://kerndatensatz-forschung.de/version1/Spezifikationstabelle_KDSF_v1.html`), there is also a table with all discussed elements (`http://kerndatensatz-forschung.de/version1/Spezifikationstabelle_KDSF_v1_komplett.html`) which includes the elements that have been rejected or postponed.

detailed definitions as given in the table view. Also, the information about rejected and postponed elements was very useful to understand what should be explicitly be *excluded* from the data model.

2.3. Conceptual Modelling

The conceptual modelling was done using the Web Ontology Language (OWL). We did not exploit the full expressive power of the language, we focused on the modelling of classes, their attributes and relationships. The main advantage of this language is that it has been standardised by the World Wide Web Consortium (W3C), and that there is a broad range of tools available supporting the work with ontologies defined in this language. For example, we used Protégé for editing the ontologies.

The resulting model is semantically very close to the textual definitions. Often, there are one-to-one relationships between elements of the data model and the corresponding definition of an element. This simplifies also the quality control of the model, i.e., the check for completeness and correctness. The completeness check can be done automatically, i.e., checking whether for each definition represented in the table, there is a corresponding ontology element. The link between ontology elements and table entries has been realised with annotations in the ontology.

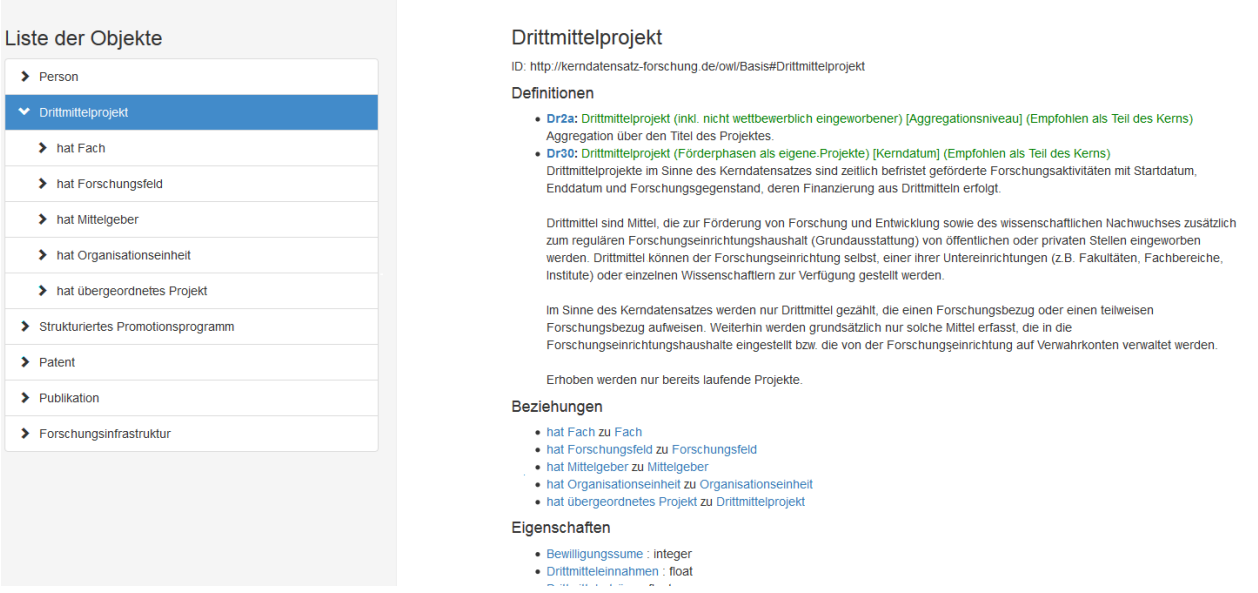

Fig. 1. Web-based representation of the conceptual model

The ontology is visualised in a web-based platform to allow also the non-expert to browse and explore the model[c]. Figure 1 shows a screenshot of the web-based documentation platform for the conceptual data model. Technically, the OWL ontology is transformed into a JSON document (JavaScript Object Notation) which is parsed by the web browser[d]. Then, the web browser creates the interactive web pages that enables the user to browse through the conceptual model. The textual definitions are directly integrated into web pages describing the ontology. A link to the tabular representation of the corresponding definition is also provided.

2.4. Mapping to CERIF

The final step in the modelling process is the mapping to the CERIF standard. We first tried to map the elements of the conceptual model directly to corresponding elements in the CERIF data model. The mapping was only possible

[c] `http://kerndatensatz-forschung.de/version1/technisches_datenmodell/`
[d] The source code for the OWL-to-JSON transformation is available at `https://github.com/chquix/OwlJsonDataModel`.

for a small part of the conceptual model that covered the more generic entities. At the detail level, a mapping was frequently not possible as CERIF simply does not cover the details of the German science system.

Thus, compatibility between the German research core dataset and CERIF could not be achieved by a simple mapping between two data models. Thus, we took a different approach to link the data models. As CERIF is defined also as an XML schema, extending the model is easily possible. Therefore, an XML schema for the German research core dataset is defined as extension of the XML Schema of CERIF.

It extends specifically some types in the schema definition, thereby enabling as much as possible the reuse of data models which are already compatible with CERIF. The link to the previously defined conceptual model and the definitions are provided as annotations of elements in the XML schema. These links are very important because they provide the information for the correct semantic interpretation of the XML schema. If this information would not be present, the schema would just define a syntactical structure. It also defines the context of the research information to be described by the XML schema.

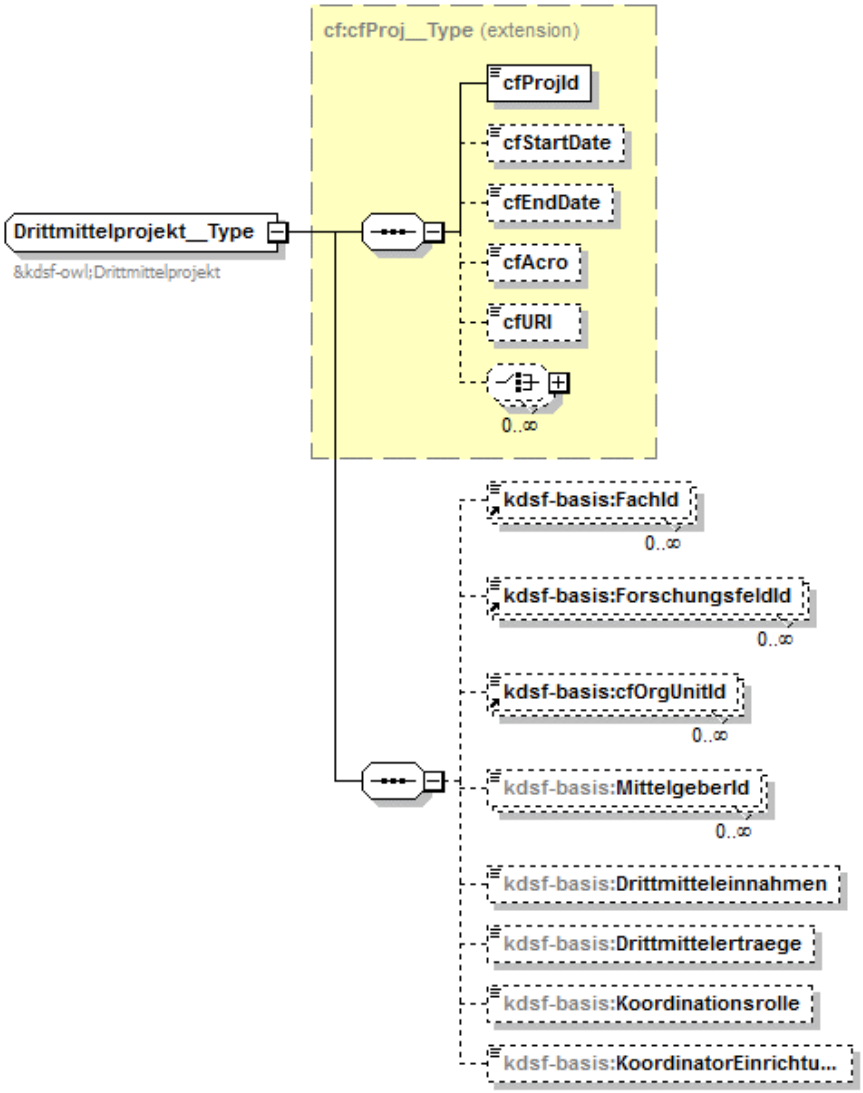

Fig. 2. Example extension of CERIF XML

Figure 2 shows an example for the extension of CERIF XML elements for the German research core dataset[e]. CERIF defines a type called `cfProj__Type` representing basic information about projects. The research core dataset required the definition for an entity type called `Drittmittelprojekt` (projects funded by third-parties), which is defined as an extension of the CERIF type. In addition to the basic attributes represented in CERIF, the element in the research core dataset should provide information about the funding agency, the amount of money received and spent, and information about the coordinator of the project. These elements have been defined as additional elements of the extended type `Drittmittelprojekt__Type`. The annotation `&kdsf-owl;Drittmittelprojekt` points to the corresponding element in the OWL ontology and thereby enables the linking of the syntactical representation of data in the XML schema to the semantical definitions in the OWL ontology.

	cfClassSchemeId	cfDescr	cfClass
1	kdsf:cfPers_Country	cfDescr cfTrans=o cfLangCode=de	cfClass (2) 1 kdsf:Staatsangehoerigkeit 2 kdsf:hatPromotionsberechtigungAus
2	kdsf:Qualifikation	cfDescr cfTrans: o cfLangCode: de Text: Klassen für Qualifikationen	cfClass (2) 1 kdsf:Habilitiert 2 kdsf:Promoviert
3	kdsf:Qualifizierungsverfahren	cfDescr cfTrans=o cfLangCode=de	cfClass (2)
4	kdsf:Faecherklassifikation	cfDescr cfTrans=o cfLangCode=de	cfClass (2)
5	kdsf:Forschungsfelder	cfDescr cfTrans=o cfLangCode=de	cfClass (2)
6	kdsf:Befristung	cfDescr cfTrans=o cfLangCode=de	cfClass (2)
7	kdsf:Finanzierungsform	cfDescr cfTrans=o cfLangCode=de	cfClass (4)
8	kdsf:Personalkategorie	cfDescr cfTrans=o cfLangCode=de	cfClass (8) 1 kdsf:SonstigesPersonal 2 kdsf:Verwaltungspersonal 3 kdsf:WissenschaftlichesUndKuenstlerischesPersonal 4 kdsf:Professoren 5 kdsf:WissenschaftlicheOderKuenstlerischeHilfskraefte 6 kdsf:WissenschaftlicheUndKuenstlerischeMitarbeiter 7 kdsf:SonstigesWissenschaftlichesUndKuenstlerischesPersonal 8 kdsf:WissenschaftsunterstuetzendesPersonal
9	kdsf:Besoldung	cfDescr cfTrans=o cfLangCode=de	cfClass (15)
10	kdsf:Professurenbezeichnung	cfDescr cfTrans=o cfLangCode=de	cfClass (3)
11	kdsf:Taetigkeitsart	cfDescr cfTrans=o cfLangCode=de	cfClass (2)
12	kdsf:Kooperation	cfDescr cfTrans=o cfLangCode=de	cfClass (4)
13	kdsf:Mittelgeber	cfDescr cfTrans=o cfLangCode=de	cfClass (11)
14	kdsf:Dokumenttyp	cfDescr cfTrans=o cfLangCode=de	cfClass (8) 1 kdsf:BibliographieDokumenttyp 2 kdsf:Editorial 3 kdsf:LetterToTheEditor 4 kdsf:MeetingAbstract 5 kdsf:QuelleneditionDokumenttyp 6 kdsf:Rezension 7 kdsf:Review 8 kdsf:WissenschaftlicherArtikel
15	kdsf:Publikationstyp	cfDescr cfTrans=o cfLangCode=de	cfClass (21)
16	kdsf:Forschungsinfrastrukturart	cfDescr cfTrans=o cfLangCode=de	cfClass (3)
17	kdsf:Forschungsinfrastrukturtyp	cfDescr cfTrans=o cfLangCode=de	cfClass (4)
18	kdsf:Zugangsart	cfDescr cfTrans=o cfLangCode=de	cfClass (3)

Fig. 3. Classificiation Schemes defined for the German Research Core Dataset

Another important feature of the CERIF standard are the classification schemes, which can be used for almost any entity or relationship in the CERIF data model. This is an easy method that enables the adaption of CERIF for country-specific classifications. We used the classification schemes to define classification for several types, for example, publication types, funding agencies, staff categories, type of funding. Figure 3 shows the defined classification schemes

[e] The modelling of the XML schema and the screenshot has been done with Altova XML Spy.

for the German research core dataset, with the details for a few classes. The first entry `kdsf:cfPers_Country`, for example, is a classification scheme for the relationship between persons and countries. CERIF just states that there is a relationship between person and countries, but does not assign a specific semantics to this relationship. For the German research core dataset, it is important to know the nationality of a person and from which country the person received the degree to participate in a PhD program. These are two different types of relationships between person and countries; thus, they are represented by two different classes in this classification scheme.

Other examples are the staff categories (`kdsf:Personalkategorie`) and document types (`kdsf:Dokumenttyp`) for which we defined the classes that have been discussed in the other working groups. These classification schemes can be easily adapted or extended if additional classes need to be considered, but it does not change the basic structure of the data model. Thus, compatibility between different versions of a dataset based on CERIF is easier to achieve.

3. Conclusion

Developing a harmonised data model for research information on a national level is a challenging task that requires a careful documentation of the whole process to reach acceptance for the model[7]. In addition, a model for research information should also address the technical challenges in research information management, such as the interoperability between research information systems.

Within the project to develop the German Research Core Dataset, we addressed those challenges and developed a tool to document and visualise the argumentation and discussion process. Additionally a data model that is compatible with CERIF was developed to enable easy exchange of research information within the context of the German science system. This context was explicitly linked to the CERIF data model by extending the CERIF XML schema with elements and classifications that are specific for Germany, but also by providing links to the intended interpretation of these elements.

Traceability of data model elements is helpful not only for the acceptance of the data model, it also provides a thorough documentation for the data modelling process. This comprehensive documentation increases the quality of the developed data models as it increases the consistency and completeness of the data model, but also decreases redundancy. Some of these quality checks have been implemented as automatic 'test' procedures that provide information to the data modeller about the completeness of his model.

The iterative nature of the discussion process has also implications for the corresponding modelling of the discussion results. Versioning and keeping track of the evolution of the data model are therefore important during the modelling process. While versioning can be easily achieved with version control systems, keeping track of the evolution of the data model is more challenging. Requirements change and the data model should be changed accordingly. However, it is sometimes difficult to identify the changes in the requirements. Instead of a new version of the requirements, a 'change log' describing the modifications to the previous version (e.g., renaming, insertion, or deletion of elements) would have been more helpful and would also provide more information to users of the data model.

The initial project to define a first version for the research core dataset has been successfully completed by the end of 2015. The maintenance and evolution of the current specifications and models will be a challenge for the next years, but the results for the documentation, visualisation, and modelling during the initial development process should be taken into account in the future.

Acknowledgements

This work has been funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project 'Kerndatensatz Forschung' (grant no. M511300C, `http://kerndatensatz-forschung.de/`).

References

1. Biesenbender, S., Hornbostel, S.. The Research Core Dataset for the German science system: challenges, processes and principles of a contested standardization project. *Scientometrics* 2016;**106**(2):837–847. URL: `http://dx.doi.org/10.1007/s11192-015-1816-y`. doi:`10.1007/s11192-015-1816-y`.

2. Riechert, M., Biesenbender, S., Quix, C.. Developing and standardising definitions for research information: Framework and methods of successful process documentation. In: *13th Intl. Conf. on Current Research Information Systems (CRIS)*. Scotland, UK; 2016, .
3. Jörg, B.. Standardising research contexts towards system interoperability - and more. In: *SK-CRIS Event*. 2014, URL: http://hdl.handle.net/11366/241.
4. Riechert, M., Dees, W.. Research information standardization as a wicked problem: Possible consequences for the standardization process. case study of the specification project of the german research core dataset. In: Jeffery, K.G., Clements, A., de Castro, P., Luzi, D., editors. *12th International Conference on Current Research Information Systems (CRIS)*; vol. 33 of *Procedia Computer Science*. Rome, Italy: Elsevier; 2014, p. 272–277. URL: http://dx.doi.org/10.1016/j.procs.2014.06.043. doi:10.1016/j.procs.2014.06.043.
5. Riechert, M., Biesenbender, S., Dees, W., Sirtes, D.. Developing definitions of research information metadata as a wicked problem? characterisation and solution by argumentation visualisation. *Program* 2016;**50**(3). URL: http://www.emeraldinsight.com/doi/abs/10.1108/PROG-01-2015-0009. doi:10.1108/PROG-01-2015-0009.
6. Conklin, J., Weil, W.. Wicked problems: naming the pain in organizations. White paper; 1998. URL: http://uuslepo.it.da.ut.ee/~maarjakr/creative/wicked.pdf.
7. Riechert, M., Quix, C., Zarnekow, R.. Fostering transparency in policy development processes - A development transparency framework. In: Becker, J., vom Brocke, J., de Marco, M., editors. *23rd European Conference on Information Systems (ECIS)*. Münster, Germany; 2015, URL: http://aisel.aisnet.org/ecis2015_rip/39.
8. Kirschner, P.A., Buckingham-Shum, S.J., Carr, C.S., editors. *Visualizing Argumentation - Software Tools for Collaborative and Educational Sense-Making*. Springer-Verl; 2003. doi:10.1007/978-1-4471-0037-9.
9. Kunz, W., Rittel, H.. Issues as elements of information systems. Tech. Rep. Vol. 131; Institute of Urban and Regional Development, University of California; 1970.
10. Quix, C., Jarke, M.. Information integration in research information systems. In: Jeffery, K.G., Clements, A., de Castro, P., Luzi, D., editors. *12th International Conference on Current Research Information Systems (CRIS)*; vol. 33 of *Procedia Computer Science*. Rome, Italy: Elsevier; 2014, p. 18–24. URL: http://dx.doi.org/10.1016/j.procs.2014.06.004. doi:10.1016/j.procs.2014.06.004.