Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 33 (2014) 18 – 24

Procedia Computer Science

CRIS 2014

Information Integration in Research Information Systems

Christoph Quix[a,*], Matthias Jarke[a,b]

[a]*Fraunhofer-Institute for Applied Information Technology FIT, 53754 St. Augustin, Germany*
[b]*RWTH Aachen University, Information Systems and Databases, 52056 Aachen, Germany*

Abstract

Information integration is an on-going challenge in data management and various approaches have been proposed in database research. New technologies and application areas create different requirements for integration systems. Research information management (RIM) is yet another challenge for data integration. RIM has many properties that are typical for data integration scenarios: many data sources, various modeling languages and data models, heterogeneity in syntax and semantics. Furthermore, many stakeholders are involved in RIM, usually with diverting goals. The combination of these properties makes RIM a particular difficult integration problem.
In this paper, we discuss the applicability of data integration approaches to research information management. In particular, we want to highlight the lessons which have been learned in data integration in the recent years. Early approaches in data integration focused on the data models and the problems with schema integration. Recent work rather concentrates on the mappings between models and integration processes. Our main argument in this paper is that mappings should be also considered as key objects in research information systems.

Keywords: research information; information integration; mappings; interoperability; standardization

1. Introduction

Management of research information has become an important issue for universities, research organizations, funding and government organizations. There is an increasing demand to collect, integrate, and analyze research

* Corresponding author. Tel.: +49-2241-14-1525.
E-mail address: christoph.quix@fit.fraunhofer.de

doi:10.1016/j.procs.2014.06.004

information for various reasons. The diverting goals of the stakeholders make research information management a difficult problem in information management. Information integration is an immanent feature of research information systems as research information has to be collected from various sources, e.g., information systems for financial budgets, libraries, human resources. Studies on research information management (e.g., UKRISS [1]) have shown that the standardization, harmonization, and integration of research information are frequently mentioned requirements and challenges, but the benefits of an integrated dataset on research information is also one of the main drivers for implementing current research information systems (CRISs). Currently, a lot of human effort is spent on collecting, integrating, and aggregating research information. As reports about the results of research are requested by various organizations, these tasks have to be done repeatedly and therefore require even more human effort. Thus, a uniform data model (or even a standard) for research information could simplify this task.

The standardization of data models for research information started already in the 1980s after first systems for the exchange of research information have been built (e.g., IDEAS [2]). As the standardization was initiated by the European Commission, the first version of CERIF (Common European Research Information Format) in 1991 focused on projects as the main entity; meanwhile, the CERIF standard has been extended and includes also information about persons, publications, organizations, etc.

Research information management (RIM) has received a growing attention in the recent years and various initiatives have been established in several countries to address issues related to research information. The reasons for RIM are manifold; one aspect is certainly collecting a core dataset to assess the quality of research. There has been and still is a lot of discussion on how this assessment can be done. We do not want to provide additional arguments for this discussion; we focus in this paper on the technical aspects of research information management and especially the standardization of a basic data model for research information.

The context of the work presented in this paper is the German project for developing a basic data model for research information[†] (called the 'core dataset', 'Kerndatensatz Forschung' in German, KDSF in short in the following), which has been initiated by the German Council of Science and Humanities (Wissenschaftsrat) and which is funded by the Federal Ministry for Education and Research. Within this project, several working groups discuss different aspects of the data model, e.g., definitions of the main elements, a classification of subjects, definitions for bibliographic data, and the definition of formal data models and interfaces. Basis for the discussion is a recommendation of the German Council of Science and Humanities in which the data model has been sketched[‡]. One goal of the project is to provide a more formal definition for the data model, including a textual definition of the semantics of the concepts and formal specification of data models which should be used to exchange research information.

One could argue that CERIF should be used as a data model for research information and the definition of another data model is not necessary. Indeed, the recommendation states that CERIF should be used as a basis for the definition of the 'core dataset', but it also specifies several data elements which are not covered by the CERIF standard. These extensions address especially aggregated data values, which can be computed from base data (e.g., number of employees, sum of all expenditures in projects funded by third-parties). Nevertheless, it is important to link the additional elements to existing elements in the CERIF standard such that the semantics of the new elements can be understood by analyzing their relationships and/or their context.

In addition, there are several other data models or standards which are relevant for the definition of a data model for research information. For example, bibliographic data is a domain in which several 'standards' are in use; publications can be identified by DOIs (Digital Object Identifiers); authors by ORCIDs (Open Researcher and Contributor ID); project information is published by different funding organizations. Thus, the development of a new data model cannot be done independently of existing information and needs to be put into the context of existing data models.

From the viewpoint of a computer scientist or data modeler, 'putting into context' means that formal relationships should be established which clearly specify how the data represented in two different data models is related. Such a

[†] http://www.forschungsinfo.de/kerndatensatz/en/

[‡] http://www.wissenschaftsrat.de/download/archiv/2855-13.pdf (in German)

relationship cannot be represented only by a 'soft' link in the sense of 'is related to'; we need to have a formal specification for a data translation method. This can be done by a mapping [3].

In this paper, we argue that CRISs should include formal mappings as first-class citizens into their data model as much of the semantics of the data is contained in the mappings. The mappings could make the management of research information more transparent as a user could verify how certain data items are computed or how his/her data is used in research reports. This could lead to a higher acceptance of CRIS by researchers because currently, many researchers ask critical questions about research information management.

The paper is structured as follows. In the next section, we will first briefly review the role of mappings in data integration and how the field of data integration research evolved in the recent years. Our approach for the application of mappings in the context of research information management in the KDSF project is discussed in section 3. Section 4 concludes our paper and points out future work.

2. Mappings in Research Information Management

CERIF provides a good basis for RIM. It addresses most of the requirements for a data model: it includes a conceptual model describing the semantics; it has well documented logical models with mappings to the conceptual models; and reference implementations in SQL and XML are also available. However, this is only a static view of research information. If we examine a certain data item (e.g., '42'), we may know what this item means (e.g., the number of journal publications of 'John Doe' in the last ten years), but we do not know how this value has been computed (e.g., 'Who did enter the publications into the system?', 'From which sources was the publication information extracted?', 'How did the system consolidate publication entries from overlapping sources?'). Therefore, it is important also to describe the process how certain data items have been derived, which data sources have been used to compute the values, and which context the data was originally created.

The formalization of a language to describe such processes would be a challenging task, especially, if also the context should be taken into account. For example, the CERIFy project [4] applied business process modeling techniques to model processes related to RIM, e.g., processes in which research information was used or produced. The goal was to show the benefit of using CERIF in such processes as the modified processes (using data in the CERIF model) should be simpler. The detailed modeling of these processes is time consuming and, because of this reason, was done only for two processes.

Data management systems focus on the 'computational', well-defined deterministic parts of such processes. A CRIS can be also seen as data management system, in which data of several other systems is going to be integrated. In the context of the German KDSF project, we interviewed several vendors as well as users of CRISs in Germany, and they agreed that a CRIS is mainly a platform for data integration. For example, within a research organization, personal data has to be extracted from a human resource system, financial data about projects is extracted from a system for the financial administration, and bibliographic data is extracted from a publication system maintained by the library of the research organization.

Thus, data integration is an integral part of CRISs. Research in data integration has evolved from schema-centric approaches to approaches which focus more on mappings. Early approaches (before 1986, [5]) dealt mainly with the aspect of schema integration, i.e., building a uniform, integrated schema for a set of source schemas. Mapping languages were rather weak as only one-to-one correspondences could be expressed, e.g., a concept is similar to another concept, or an attribute is the same as another attribute in the other schema. However, the heterogeneity of information systems and their schemata requires often a more expressive mapping language, as the relationships between different data models cannot be captured by simple links having a semantics like 'sameAs' or 'isSimilarTo'.

The first ideas for model management systems [6] considered both, models and mappings, as first-class objects. A model management system should provide formal methods for frequent operations which are required in building model-intensive applications (including integrated information systems). Such operations include the matching of models, merging of models (schema integration), mapping generation and composition. The vision for model management systems opened up a new research area in which several systems [7, 8, 9] have been developed.

However, the representation of mappings was also very weak in the beginning, e.g., informal mappings which state correspondences between model elements; such correspondences are often the result of schema matching [10]

and the semantics usually does not go beyond 'similar'. Human effort is often required to verify and formalize the output of a schema matching process. Such mappings cannot directly be applied in a data integration process. They can only support the developer in finding correspondences in different data sources.

To be really useful for data integration, a mapping must be specified in a formal language. Later approaches [3, 11, 12] in model management provided more expressive mapping languages which can be used for powerful data transformations in heterogeneous environments. These mappings can be seen as a formal specification for a data transformation process from a source database to a target database. Basically, a mapping can be seen as a pair of queries q_s and q_t which states a relationship the answers of the queries, e.g., equal or subset. The query q_s is a query over a data source, and q_t is a query over a target database. Thus, q_s specifies how the data should be extracted from the data source, and q_t specifies how the data should be inserted into the target databases. Formal mappings can be also constructed in an incremental way as a complete integration of all data sources is usually not necessary and too difficult in the early stages of a systems [13]. Such an approach has also been proposed for CRIS [14].

3. Towards an Approach for Research Information Integration

The process of building a standard data model for research information shows some parallels to the research in data integration. As stated above, research information management requires the integration of data from different sources. The development of a data model for a CRIS is therefore very similar to the process of building an integrated schema, which requires the definition of the integrated schema and the mappings to the data sources.

The approach which we propose in the German KDSF project follows this idea and specifies models as well as mappings between data models as illustrated in Fig. 1. Starting point for the development of the formal models are the textual definitions of concepts and their relationships which are created by several working groups. A technical working group has the task of translating these definitions into formal data models.

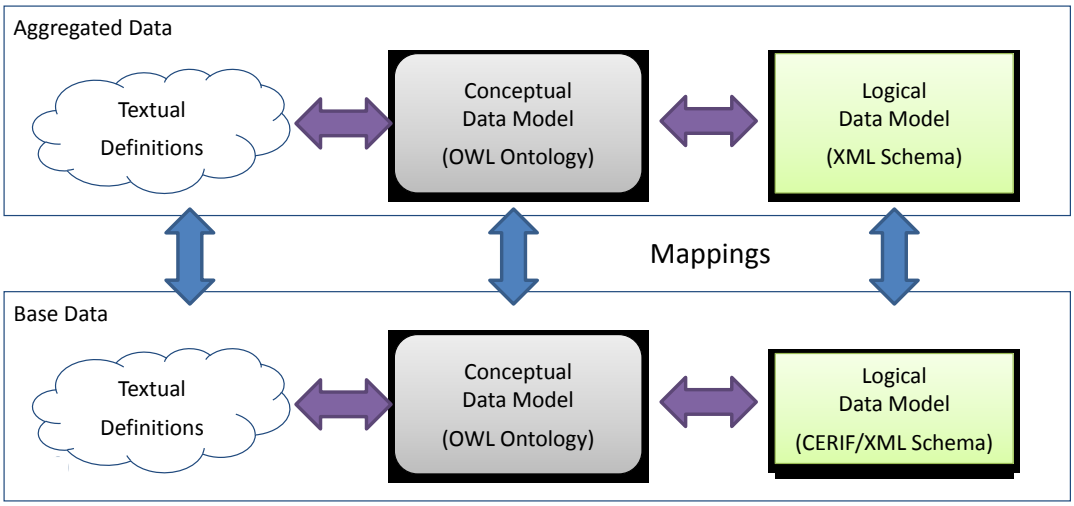

Fig. 1. Mappings between data models.

Within the definitions, we can distinguish two different types of data:

- **Base data** are individual data items describing certain data objects on a detailed level, e.g., persons, research units, publications, projects. The CERIF standard provides a good foundation for the description of this base data and it is planned to use the CERIF XML Schema in the KDSF project.
- **Aggregated data** is data that can be computed by aggregating certain base data items, e.g., the number of publications for a research unit, the number of researchers in a project.

On both levels, we will have textual definitions and the corresponding representations in the formal data models. We decided to follow the classical approach for data modeling also in this project. The textual definitions will be first translated into a conceptual model, and then a translation into a logical model will be done.

3.1. Choice of the Modeling Language

The conceptual model should focus on the definition of the concepts and their relationships. A representation of the model in a way that can be easily understood should be also possible. Candidates for conceptual modeling languages are the Entity Relationship Model (ERM) [15], the Unified Modeling Language (UML), and the Web Ontology Language (OWL). The ERM is frequently used in database design, but it is not standardized (or at least, there is no accepted standard) and many variants of the modeling language are in use. Because of this, the support by modeling tools is also limited. Thus, the ERM is not a good candidate for the definition of the conceptual model.

UML and OWL are both languages that have been standardized, are used in various application domains and for which several modeling tools are available. However, UML covers much more aspects than just data modeling and is therefore a more complex language. The interoperability of models developed with different tools is also frequently a problem in UML.

Conceptual models can be also specified as ontologies in OWL, which is also a complex language as it is derived from description logics [16]. However, most ontologies (including the ontologies which we will define in the KDSF project) use only a simple subset of the language (known as RDF Schema) in which concepts, their properties, and their relationships (including subclass relationships) can be specified. In this case, a representation of an ontology as a graph is also possible. As OWL has been developed for the Semantic Web [17], linking of ontologies to other ontologies is an immanent feature of the language. Therefore, we have chosen OWL as the language for the conceptual data model.

On the logical level, there is not really an alternative to XML. One goal of the KDSF project is to enable the exchange of research information. XML has been established as the data format for data exchange between different systems. Furthermore, CERIF has also a definition as XML Schema and this format is already supported by some CRIS implementations. Thus, the KDSF project will also use XML Schema to define the data exchange format in which the existing CERIF XML Schema will be integrated.

3.2. Mappings for Aggregated Data

During the project, the textual definitions and the definitions of the data models will be continuously developed. It is therefore necessary to synchronize the three different definitions of the data model (textual, conceptual, and logical). As OWL and XML are web languages, they both provide methods to annotate elements with additional metadata. This metadata can be a URI (Uniform Resource Identifier) that identifies the corresponding element in another data model. For example, if the XML Schema defines an element, its metadata can refer to the corresponding element in the ontology, and an element in the ontology has metadata which refers to the textual definition of the concept. Therefore, it is necessary that all elements in the data models are identified by URIs.

By doing so, we can also establish basic checks for the completeness and correctness of the data models. For example, we could verify whether each element in the textual definition has a corresponding element in the conceptual data model, i.e., completeness of the conceptual data model. For example, the correctness of the logical model can be verified whether it contains any elements which do not have a corresponding element in the conceptual model. These checks can be even automated and are especially important during the evolution of the data models. In a new version of a data model, several elements might be added or removed; therefore, it is important to apply the corresponding changes also in the other data models.

These mappings can be realized with the annotation features of OWL and XML and are represented by the horizontal links indicated in Fig. 1. In addition, the vertical relationships between the base data and the aggregated data need to be represented as well. However, in this case, we have to use a more complex formalism as we do not have any more simple one-to-one correspondences as in the horizontal mappings, e.g., a set of data items is aggregated into one data value. To avoid the definition of a new formalism for this purpose, we propose to reuse existing query languages for the definition of the aggregated data items.

SPARQL[§] and XQuery[**] are query languages for OWL and XML, respectively. Both are very expressive languages, syntax and semantics have similarities with SQL. Thus, these languages are sufficient for the purpose of defining the mappings between base data and aggregated data. The following query shows an example for a SPARQL query computing the number of employees for organizational units:

```
SELECT ?Subject ?OrgUnit (count(?Employee) AS ?CountEmpl)
{     ?Employee   :hasEmployment    ?Employment
      ?Employment :hasSubject       ?Subject
                  :employingUnit    ?OrgUnit
      ?OrgUnit    :hasSubject       ?Subject }
GroupBy ?Subject, ?OrgUnit
```

Such a query can be attached as annotation to a property definition in the conceptual data model. A similar query can be also specified in XQuery and could be annotated to the corresponding element in the XML schema. Note that this kind of annotations should be understood as additional information for the interpretation of certain data elements and not as a mandatory way for implementing the data model. The same applies to the choice of the modeling and query languages. OWL, XML, SPARQL, and XQuery have been chosen because they can be easily used to express relationships between different data models. The implementation of a CRIS could still use internally a different data model (in fact, CRISs usually rely on the relational model), just for data exchange the data should be transformed into the required XML format.

The query languages SPARQL and XQuery are also designed for the web and enable queries across several data sources (e.g., XML documents). The approach, which we have described for the mappings between base data and aggregated data, could be applied as well to relate data from different sources. It is important that the mappings between data sources and the integrated system (e.g., a CRIS) are made explicit and that they are not hidden inside a data transformation script.

4. Conclusion and Future Work

The goal of the paper was to discuss the results of data integration research in the context of research information management and to state a vision for further standardization efforts of research information. We emphasized that one important role of CRISs is data integration and that mappings are an important part of a data integration system. By using mappings, we can provide formal specifications for the implementation of a CRIS, e.g., how should be the data transformed and integrated. Furthermore, the mappings can provide lineage information of aggregated data items in research reports which supports the reproducibility, traceability, and transparency of aggregated values.

This is work in progress and the German KDSF project has the ambitious goal to provide definitions and data models for several sub-areas of research information. In addition to the textual definitions of the concepts of a core dataset [18], the technical implementation of such definitions is a challenge. Initial versions of the conceptual and logical data models have been developed, but definitions and data models are continuously improved, extended and revised. The synchronization of these artefacts is a key requirement in the standardization process to avoid inconsistencies. The implementation of an adequate version management is therefore necessary; the mappings have also to be maintained between different versions of the same data model. Furthermore, the initial models have a high complexity (number of elements and their relationships) such that the visualization and representation in an understandable way becomes a challenge. This is especially important in this setting as many stakeholders are not computer scientists, but their feedback is necessary for the standardization process. Thus, creating appropriate visual representations of the data models is another challenge that has to be addressed in the near future.

[§] http://www.w3.org/TR/sparql11-overview/

[**] http://www.w3.org/TR/xquery-30/

Acknowledgements & Disclaimer

This work has been supported by and has been done in the context of the German project for developing a core dataset for research information (http://www.forschungsinfo.de/kerndatensatz/en/). Although the authors are involved in a working group of this project, the paper does not represent a formal output of the project.

References

[1] S. Waddington, A. Sudlow, K. Walshe, R. Scoble, L. Mitchell, R. Jones, S. Trowell, Feasibility study into the reporting of research information at a national level within the uk higher education sector, New Review of Information Networking 18 (2) (2013) 74–105. doi: 10.1080/13614576.2013.841446

[2] K.G.Jeffery, J. O. L.-F. Miquel, S. Zardan, F. Naldi, I. V. Parenti, Ideas: A system for international data exchange and access for science, Information Processing and Management 25 (6) (1989) 703–711.

[3] P. A. Bernstein, S. Melnik, Model management 2.0: Manipulating richer mappings, in: Proc. SIGMOD, Beijing, China, 2007, pp. 1–12. doi:http://doi.acm.org/10.1145/1247480.1247482.

[4] M. Mahey, N. Brennan, CERIFy: Using Business Process Mapping to Engage with Research Information Management Processes and the CERIF Data Model, in: K. G. Jeffery, J. Dvorak (Eds.), E-Infrastructures for Research and Innovation: Linking Information Systems to Improve Scientific Knowledge Production: Proceedings of the 11th International Conference on Current Research Information Systems, Prague, Czech Republic, 2012, pp. 147–156.

[5] C. Batini, M. Lenzerini, S. B. Navathe, A comparative analysis of methodologies for database schema integration, ACM Computing Surveys 18 (4) (1986) 323–364.

[6] P. A. Bernstein, A. Y. Halevy, R. Pottinger, A vision for management of complex models, SIGMOD Record 29 (4) (2000) 55–63.

[7] S. Melnik, E. Rahm, P. A. Bernstein, Developing metadata-intensive applications with rondo, Journal of Web Semantics 1 (1) (2003) 47–74.

[8] L. M. Haas, M. A. Hernández, H. Ho, L. Popa, M. Roth, Clio grows up: from research prototype to industrial tool, in: F. Ozcan (Ed.), Proceedings of the ACM SIGMOD International Conference on Management of Data, ACM, Baltimore, Maryland, USA, 2005, pp. 805–810.

[9] D. Kensche, C. Quix, X. Li, Y. Li, GeRoMeSuite: A system for holistic generic model management, in: Proc. VLDB, 2007, pp. 1322–1325.

[10] E. Rahm, P. A. Bernstein, A survey of approaches to automatic schema matching, VLDB Journal 10 (4) (2001) 334–350.

[11] R. Fagin, L. M. Haas, M. A. Hern´andez, R. J. Miller, L. Popa, Y. Velegrakis, Clio: Schema mapping creation and data exchange, in: A. Borgida, V. K. Chaudhri, P. Giorgini, E. S. K. Yu (Eds.), Conceptual Modeling: Foundations and Applications, Vol. 5600 of Lecture Notes in Computer Science, Springer, 2009, pp. 198–236.

[12] D. Kensche, C. Quix, X. Li, Y. Li, M. Jarke, Generic schema mappings for composition and query answering, Data Knowl. Eng. 68 (7) (2009) 599–621. doi:http://dx.doi.org/10.1016/j.datak.2009.02.006.

[13] M. Franklin, A. Halevy, D. Maier, From databases to dataspaces: a new abstraction for information management, SIGMOD Record 34 (4) (2005) 27–33. doi:http://doi.acm.org/10.1145/1107499.1107502.

[14] K. Jeffery, A. Asserson, CRIS and DataSpaces by Epitaxial Growth, in: Connecting Science with Society -The Role of Research Information in a Knowledge-Based Society: Proceedings of the 10th International Conference on Current Research Information Systems, Aalborg, Denmark, 2010.

[15] P. P. Chen, The entity-relationship model -toward a unified view of data, ACM Trans. Database Syst. 1 (1) (1976) 9–36.

[16] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, P. F. Patel-Schneider (Eds.), The Description Logic Handbook: Theory, Implementation, and Applications, Cambridge University Press, 2003.

[17] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Scientific American 284 (5) (2001) 34–43.

[18] M. Riechert, W. Dees, 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. Proceedings of the 12th International Conference on Current Research Information Systems (CRIS 2014), Rome, Italy, 2014.