CERIF: THE COMMON EUROPEAN RESEARCH INFORMATION FORMAT MODEL

Brigitte Jörg

German Research Center for Artificial Intelligence (DFKI), Language Technology Lab, Berlin, Germany
Email: brigitte.joerg@dfki.de

ABSTRACT

With increased computing power more data than ever are being and will be produced, stored and (re-) used. Data are collected in databases, computed and annotated, or transformed by specific tools. The knowledge from data is documented in research publications, reports, presentations, or other types of files. The management of data and knowledge is difficult, and even more complicated is their re-use, exchange, or integration. To allow for quality analysis or integration across data sets and to ensure access to scientific knowledge, additional information – Research Information – has to be assigned to data and knowledge entities. We present the metadata model CERIF to add information to entities such as Publication, Project, Organisation, Person, Product, Patent, Service, Equipment, and Facility and to manage the semantically enhanced relationships between these entities in a formalized way. CERIF has been released as an EC Recommendation to European Member States in 2000. Here, we refer to the latest version CERIF 2008 – 1.0.

1 INTRODUCTION

Research is becoming more and more data intensive. With increased computing power more data than ever are being and will be produced, stored and (re-) used. Data are collected in databases, computed and annotated, or transformed by specific tools. The knowledge from data is documented in research publications, reports, presentations, or other types of files. The management of data and knowledge is difficult, and even more complicated is their re-use, exchange, and integration. To allow for quality analysis or integration across data sets and to ensure access to scientific knowledge, additional information – Research Information – has to be assigned to data and knowledge entities. Especially with recent developments in national assessment and performance exercises, Research Information as an asset is gaining ground. The applied evaluation methods depend upon formalized information, and information quality becomes a critical issue (Asserson & Simons, 2006; Bosnjak & Stempfhuber, 2008). Not only at a national level but also at the European scale, Research Information is being recognized as a player alongside publication repositories to improve the access to scientific knowledge (Driver, 2008) and as an enabler for large-scale data integration and data management (Joint, 2008; Carpenter, 2008). Most European countries collect and store their Research Information in digital repositories; these may be national, regional, institutional, functional, or thematic in their range, where each system builds upon a particular format or structure to serve for special requests. In order to gain additional value from data and knowledge distributed across systems, the information assigned has to be integrated. That is, the individual information structures and information system formats have to be mapped towards an agreed format within a target system for further analysis and access. Information integration is not an easy task, difficult at the national level and quite a challenge at the European scale (i.e., Jörg et al., 2008) or beyond. However, analysis of and access to scientific data, knowledge, and the information assigned, is an essential requirement in the ERA[1], for innovators, academics, decision makers, media, and the members of the society in general. It is realized that research and development leads to wealth creation and improvement in the quality of life. Because public funding is involved, it is necessary for there to be appropriate governance and also for the information to be available to the public.

CRIS and CERIF approaches into this direction are not new (Asserson et al., 2002). In the 1970s serious efforts for international cooperation among research information systems were made to survey a country's scientific and

1 European Research Area (ERA): http://ec.europa.eu/research/era/index_en.html

technological potential and to use such information in the formulation of the science policy on a national level[2]. In 1971, Unisist[3] published a "Study report on the feasibility of a world science information system" (Unisist, 1971). In 1987 the European Working Group on Research Databases held a workshop and, as a result, recommended CERIF to be used as a standard format to permit exchange of records among different European member countries and to serve as a basis for setting up a network among research databases.

Each nation state has similar research processes: strategic planning; program announcement; call for proposals; proposal evaluation and awarding; project result monitoring; and project result exploitation. However, research is international. A research project in one country is likely based on previous research in several other countries. Many research projects are transnational. Knowledge about the research activity in one country may influence the strategy towards the research, including priorities and resources provided, in another country. Thus, there is a need to share such information across countries or even between different funding agencies in the same country. Research Information is used by researchers (to find partners, to track competitors, to form collaborations); research managers (to assess performance and research outputs and to find reviewers for research proposals); research strategists (to decide on priorities and resourcing compared with other countries); publication editors (to find reviewers and potential authors); intermediaries/brokers (to find research products and ideas that can be carried forward with knowledge/technology transfer to wealth creation); the media (to communicate results of R&D in a socio-economic context); and the general public (for interest). Research Information is relevant for actors in scientific environments as well as for decision makers to support related organization, management and planning. We consider Research Information as the transmitter between Science and Society and, as such, as a powerful instrument for governance. Having such an impact, Research Information has to be collected carefully and preserved systematically, in order to most effectively support society and the individuals within (EuroHORCS, 2008).

2 CURRENT RESEARCH INFORMATION SYSTEMS (CRISs)

Research Information is managed in research information systems. They allow for a coherent view over information about research actors, their activities and their environments (Jeffery & Asserson, 2006a).

Research Information Systems are built upon conceptual domain models to capture the meaning of the domain by structuring it into entities and their relationships (Wand & Weber, 2002). As entities we consider the objects, such as Person, Project, Organization, Publication, Patent, Product, Funding, Equipment, and Facility, relevant in the Research domain. An entity can be represented by attributes and by the relationships it maintains with other entities at a time. The relevant entities, their attribute and relationship descriptions as such, compose the model of the domain for setting up a particular information system. In the CRIS community, we preferably talk of *Current* Research Information Systems (CRISs) to indicate their dynamics and timeliness (Jeffery & Asserson, 2006b). Some example questions that may be answered from a CRIS are:

- Which related project exists within the research group or organization or scientific network researcher X is part of?
- By which funding agencies or sponsors is research project A financed?
- How often have articles by author X been cited?
- Did author X publish with institutionally external authors?
- In how many FP7 projects does organization Z participate?
- How many publications have resulted from project Y?
- How many women have been involved in FP5 or FP6 projects?

---

2 CORDIS comprehensive information about CERIF, CRISs and their history: http://cordis.europa.eu/cerif/

3 UNISIST: Unesco's World Scientific Information Programme

3 THE COMMON EUROPEAN RESEARCH INFORMATION FORMAT (CERIF)

CRIS activities and developments in Europe are tightly interrelated with CERIF. CERIF is considered a standard recommended by the European Union to its Member States[4]. The physical CERIF model is a relational database model available as SQL scripts based on common ERM (Entity Relationship Model) constructs (Chen, 1976). The latest releases include a formalized, so called "Semantic Layer," and an XML interchange format (Jörg et al., 2009b).

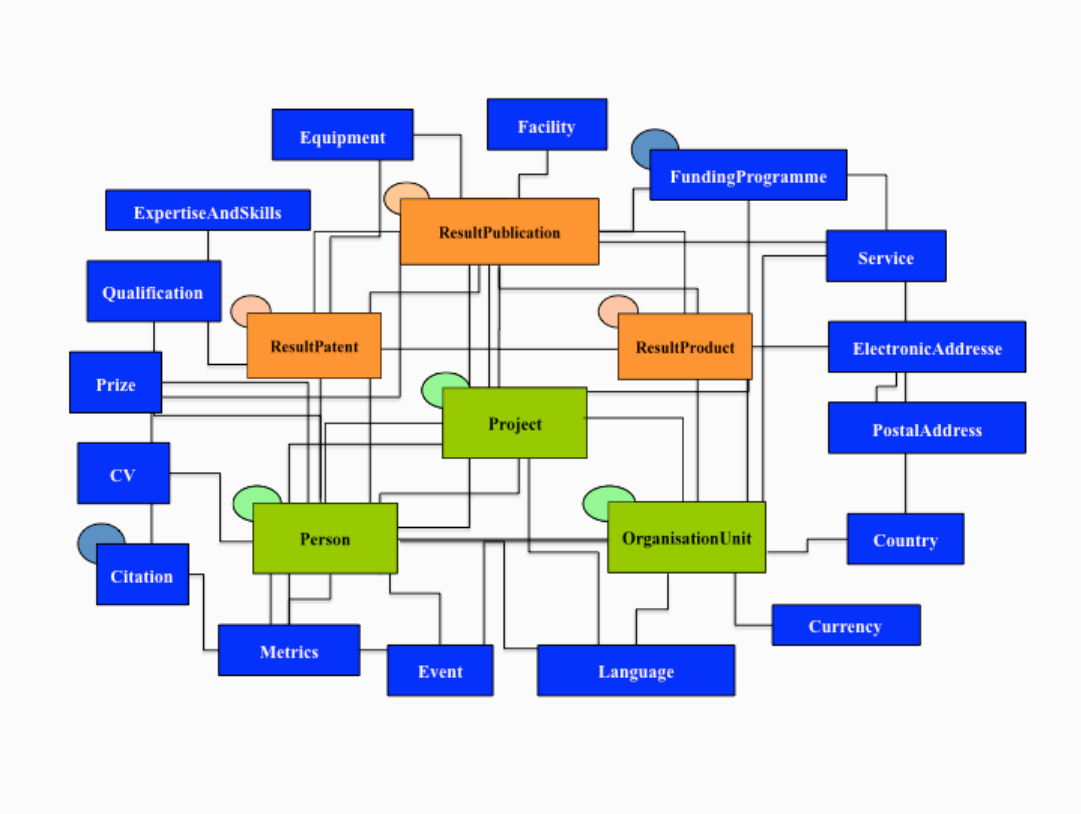

Figure 1. *CERIF Entities and their Relationships*

Figure 1 shows the CERIF entities considered relevant to represent the research domain and some of the relationships among them.

3.1 Conceptual CERIF Entity Types and Features

The CERIF model is conceptually structured into entity types and features. Among the types, we distinguish core, result, link, and 2nd level entities. As features, we consider multilinguality and semantics.

- **CERIF Core Entities (core):** The core entities are Person, OrganisationUnit, and Project. They allow for the representation of scientific actors. Figure 1 shows them in the bottom center, indicating their recursive (circles) and linking relationships. Each core entity links to itself and maintains relationships with many other entities.

- **CERIF Result Entities (result):** The result entities are ResultPublication, ResultPatent, and ResultProduct. They allow for the representation of research output. Figure 1 shows them in the upper center, indicating their relationships.

[4] CERIF: http://cordis.europa.eu/cerif/

- **CERIF 2nd Level Entities (2nd):** The 2nd level entities are, i.e., Funding, Facility, Equipment, Prize, CV, Expertise, Qualification, Citation, Metrics, Event, PostalAddress, and ElectronicAddress. They allow for the representation of the research environment. Figure 1 shows the 2nd level entities surrounding the core and result entities.

- **CERIF Link Entities (link):** The link entities are considered a major strength of the CERIF model. Link entities are the reified relationships between core, result, and 2nd level entities. A link entity always connects two entities and includes a time-stamped reference to a classifier that is itself assigned to a classification scheme.

- **CERIF Multilingual Features (lang):** CERIF supports multiple language features for names, titles, descriptions, keywords, abstracts, and even for the semantics.

- **CERIF Semantic Features (class):** The so called CERIF Semantic Layer is considered a container to manage and maintain the formal semantics (contexts) as established with the link entities. It allows for the representation of relationship kinds (Storey 1993; Wang et al 1999), application views, subject headings, any classification scheme, or even the mapping among schemes.

The presented conceptual structure of CERIF types and features is only a virtual structure and, as such, not inherent in the physical data model. It is meant to support the understanding of the model and follows the CERIF 2008 – 1.0: Model Introduction and Specification document (Jörg et. al. 2009a) to which we refer for more details.

3.2 CERIF Modularity and Components

We have presented the CERIF entity types as well as their multilingual and semantic features to demonstrate the range of coverage and the flexibility of the model with respect to research contexts. The CERIF model aims to represent the research domain, and due to its modularized and consistent structure, it allows for a selection of sub domains or 'components' with respect to particular application contexts and requirements as indicated in Figure 2.

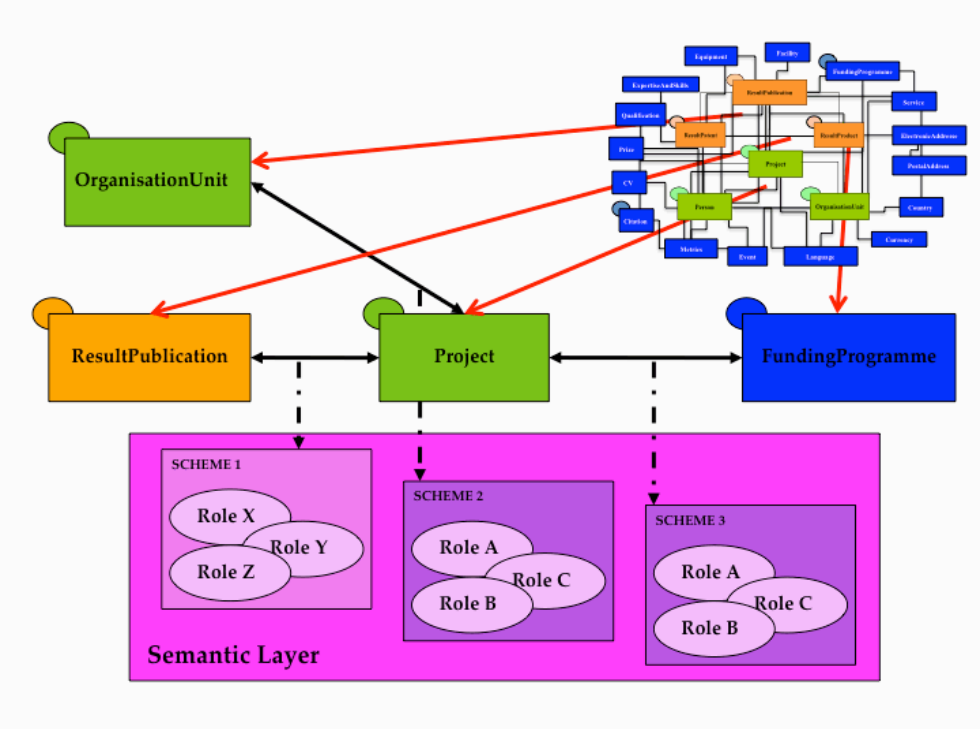

Figure 2. *Some selected CERIF entities, indicating the management of their semantics within links (simplified view)*

3.3 CERIF Example Records

The two tables below show examples of CERIF-driven database records: the first representing a Project in context, the second representing a Publication in context.

Table 1. *CERIF Project Example Record*

CERIF Project example database entry				**Semantic Layer (CERIF Semantics)**	
Data	**Entity/Table**	**Type**	**Attribute**	**Classification**	**Classification Scheme**
project-ist-world	**cfProj**	**core**	**cfProjId**		
IST World	**cfProj**	**core**	**cfAcro**		
http://www.ist-world.org/	**cfProj**	**core**	**cfURI**		
2005-04-01	**cfProj**	**core**	**cfStartDate**		
2007-11-30	**cfProj**	**core**	**cfEndDate**		
Knowledge Base for RTD Competencies in IST	**cfProjTitle**	**lang(en,o)**	**cfTitle**		
Wissensbasis für RTD Kompetenzen im Bereich IST	**cfProjTitle**	**lang(de,h)**	**cfTitle**		
IST, Research Information, NMS, Portal, Information System	**cfProjKeyw**	**lang(en.o)**	**cfKeyw**		
The objective of the project is to set up and populate an information portal …	**cfProjAbstr**	**lang(en.o)**	**cfAbstr**		
classification-2004-ist-3	**cfProj_Class**	**link**	**cfClassId**	**2004-IST-3**	**FP6-IST**
publication-analyzing-european-research-competencies-in-ist	**cfProj_ResPubl**	**link**	**cfResPublId**	**is originator of**	**PROJ-PUBL**
publication-cris-information-systems-for-research-activity	**cfProj_ResPubl**	**link**	**cfResPublId**	**is originator of**	**PROJ-PUBL**
publication-analytic-services-for-the-era-publication	**cfProj_ResPubl**	**link**	**cfResPublId**	**is originator of**	**PROJ-PUBL**
organisation-dfki	**cfProj_OrgUnit**	**link**	**cfOrgUnitId**	**is coordinated by**	**PROJ-ORG**
funding-programme-fp6	**cfProj_Fund**	**link**	**cfFundId**	**is funded by**	**PROJ-FUND**

Table 1 represents a CERIF project record where the common (core) and multilingual (lang) attributes are stored in upper rows. The lower rows show some releationships (link), including their formalized contextual semantics. Linkage is physically established by ids (cfClassId, cfResPublId, cfOrgUnitId, cfFundProgId) as indicated in the Attribute column. The Type column indicates the conceptual entity type (core, link, lang); the formalized semantic values ("2004-IST-3","is originator of", "is coordintated by", "is funded by") are stored in the Classification column, which belongs to the Semantic Layer, where each value is assigned to a predefined scheme ("FP6-IST","PROJ-PUBL","PROJ-ORG","PROJ-FUND").

In the same way, Table 2 represents a CERIF publication record where the common (result) and multilingual attributes (lang) are stored in the upper rows. The lower rows again show some relationships (link), including their formalized contextual semantics. The physical linkage is again established by ids (cfClassId, cfResPublId2, cfPersId, cfOrgUnitId, cfProjId, cfEventId), as indicated in the Attribute column. The Type column indicates the conceptual entity type. The formal semantic values ("Conference Proceedings Article", "is part of", "is author 1 of", "is publisher of", "is originator of", "is presented at") are stored in the Classification column, where each value again belongs to a predefined scheme ("CERIF2008-RESPUBL-TYPES", "RESPUBL-RESPUBL-ROLES", "PERS-RESPUBL-ROLES", "ORG UNIT-RESPUBL-ROLES").

***Table 2**: CERIF ResultPublication Example Record*

CERIF ResultPublication example database entry				**Semantic Layer (CERIF Semantics)**	
Data	**Table**	**Type**	**Attribute**	**Classification (ClassIds)**	**Classification Scheme**
publication-joerg-et-al	**cfResPubl**	**result**	**cfResPublId**		
2008	**cfResPubl**	**result**	**cfResPublDate**		
107	**cfResPubl**	**result**	**cfStartPage**		
123	**cfResPubl**	**result**	**cfEndPage**		
978-961-6133-38-8	**cfResPubl**	**result**	**cfISBN**		
http://www.eurocris.org/fileadmin/Upload/Events/Conferences/CRIS2008/Papers/cris2008_Joerg.pdf	**cfResPubl**	**result**	**cfURI**		
Analyzing European Research Competencies	**cfResPublTitle**	**lang(en.o)**	**cfTitle**		
Results from a European SSA Project	**cfResPublSubtitle**	**lang(en.o)**	**cfSubtitle**		
With this paper we will present the approach of analyzing research …	**cfResPublAbstr**	**lang(en.o)**	**cfAbstr**		
IST, ERA, CRIS, CERIF, Research Competencies, Analysis, Visualization	**cfResPublKeyw**	**lang(en.o)**	**cfKeyw**		
classification-conf-proc-article	**cfResPubl_Class**	**link**	**cfClassId**	**Conference Proceedings Article**	**CERIF2008-RESPUBL-TYPES**
publication-get-the-good-cris-going	**cfResPubl_ResPubl**	**link**	**cfResPublId2**	**is part of**	**RESPUBL- RESPUBL-ROLES**
person-brigitte-joerg	**cfPers_ResPubl**	**link**	**cfPersId**	**is author 1 of**	**PERS-RESPUBL-ROLES**
person-hans-uszkoreit	**cfPers_ResPubl**	**link**	**cfPersId**	**is author of**	**PERS-RESPUBL-ROLES**
person-jure-ferlez	**cfPers_ResPubl**	**link**	**cfPersId**	**is author of**	**PERS-RESPUBL-ROLES**
person-mitja-jermol	**cfPers_ResPubl**	**link**	**cfPersId**	**is author of**	**PERS-RESPUBL-ROLES**
orgunit-izum	**cfOrgUnit_ResPubl**	**link**	**cfOrgUnitId**	**is publisher of**	**ORGUNIT-RESPUBL-ROLES**
project-ist-world	**cfProj_ResPubl**	**link**	**cfProjId**	**is originator of**	**PROJ-RESPUBL-ROLES**
event-cris-2008	**cfResPubl_Event**	**link**	**cfEventId**	**is presented at**	**RESPUBL-EVENT-ROLES**

From the two examples, it becomes clear that each CERIF entity record is composed from different entity types and features. The separation of link entities from core, result, and second level entities allows for a rich flexibility with respect to semantic coverage, information integration, and thus applications.

4 RELATED ACTIVITIES

A survey about standards and formats in the digital library community revealed that there are many different schemes (standards) available in the library domain. Each scheme was singularly developed and not designed as an overall architecture to cover integrated object entities. For interoperability and networking in the digital age, the issues of duplicate information, overlap in sections of metadata, need rules that are currently being addressed by good practise guidelines. The resulting report recommends overcoming the problem by best practice guidelines and by pragmatic applications. The report proposes to structure metadata into:

- **Descriptive:** intellectual content
- **Administrative:** technical (file formats), rights management, provenance (creation, subsequent treatment, responsibility, ...)
- **Structural:** internal structure of items (page, order, ...)

With the survey, it was recognized that a combination of metadata standards will always be messier than the utilization of a single standard to combine taxonomic powers and to resolve potential clashes or duplications among them. Furthermore, the report revealed that integration by itself would be of little consequence if a common standard fails to address the metadata needs of the digital library community (Gartner, 2008).

CERIF allows for the representation of different standards and structures and, at the same time, enables their integration and mapping towards a common format.

5 CONCLUSION

The results from the above survey within the library community show that there is increased need for an overarching format to enable quality data integration and interoperability. An overarching standard is advantageous not only for information management but furthermore for advanced data analysis and to grant access to the data, information, and knowledge. The CERIF format offers a model to structure the research domain into relevant objects and their relationships. Moreover, with the Semantic Layer it provides a powerful means for the management of contextual semantics. The current interest, usage, and applications of the CRIS concept and the CERIF model and interchange format encourage further developments. The latest release incorporates a formalization of Publication types and Publication-related links (Jörg et al., 2009c). The priority for formalizing further contexts, i.e., for Funding or Patents will again emerge from ongoing community and task group activities.

6 ACKNOWLEDGEMENTS

I wish to thank DFKI for the support to enable my activities within euroCRIS. Additionally, I wish to thank the CERIF task group members for active participation and lively in discussions, in particular for the communication with respect to the CERIF Semantics. The work presented is also supported by the German Federal Ministry of Education and Research (BMBF) within the TAKE project under the contract 01IW08003. For some years, the work was supported by the European Commission through the project IST World under the contract FP6-2004-IST-3 – 015823).

7 REFERENCES

Bosnjak A. & Stempfhuber, M.(eds.) (2008) Get the Good CRIS Going: Ensuring Quality of Service for the User in the ERA. *9th International Conference on Current Research Information Systems*. Maribor, Institute of Information Science, June 2008.

Asserson, a. & Simons, E. J. (eds.) (2006) Enabling Interaction and Quality: Beyond the Hanseatic League. *8th International Conference on Current Research Information Systems.* Bergen, Norway, May 2006. Leuven, Leuven University Press.

Asserson, A., Jeffery, K.G., & Lopatenko, A (2002) CERIF: Past, Present and Future: An Overview. In *Proceedings: Gaining Insight from Research Information. 6th International Conference on Current Research Information Systems*, Kassel, Germany.

Carpenter, N. (2008) Tune it up: Creating and Maintaining the Institutional Repository Revolution. *Open and Libraries Class Journal,* Vol. 1, No. 1..

Chen, P.P. (1976) The entity-relationship model: Toward a unified view of data. *ACM Trans. Database Syst.* 1, 1, 9-36.

EuroHORCS (2008) Window to Science: Information Systems of the European Research Organisations. *Report of the EUROHORCS – ESF Working Group on a Joint Research Information System.* Strasbourg: European Science Foundation. Report Editing: Alexis-Michel Mugabushaka. ISBN: 2-912049-86-5, October 2008.

Gartner, R. (2008) Metadata for the digital libraries: state of the art and future directions (1.0). Peer reviewed report from the *JISC Technology and Standards Watch*. April, 2008, Bristol, UK.

Joint, N. (2008) Current research information systems, open access repositories and libraries. *Library Review. 2008*, Vol. 57, pp 570-575, Emerald Group Publishing Limited, ISSN: 0024-2535.

Jeffery, K. & Asserson, A. (2006a) CRIS: Central Relating Information System. In: Asserson A. & Simons, E. (Eds). Enabling Interaction beyond the Hanseatic League. *8th International Conference on Current Research Information Systems,* May 2006, Bergen, Norway, Leuven: Leuven University Press, 109-119.

Jeffery, K. & Asserson, A. (2006b) Supporting the Research Process with a CRIS. In: Asserson A. & Simons, E. (Eds). Enabling Interaction beyond the Hanseatic League. *8th International Conference on Current Reeearch Information Systems*, May 2006, Bergen, Norway, Leuven: Leuven University Press, 109-119.

Jeffery, K.G., Asserson, A., & Lopatenko, A. (2002). Comparative Study of Metadata for Scientific Information: The place of CERIF in CRISs and Scientific Repositories. *Gaining Insight from Research Information. 6th International Conference on Current Research Information Systems*, Kassel, Germany.

Jörg, B., Ferlez, J., Uszkoreit, H., &, Jermol, M. (2008). Analyzing European Research Competencies in IST: Results from a European SSA Project. In *Proceedings: 8th International Conference on Current Research Information Systems*. Maribor, June 2008.

Jörg, B., Jeffery, K., Asserson, A., & van Grootel, G. (2009) CERIF 2008 – 1.0 Full Data Model (FDM). *Introduction and Specification. euroCRIS*, April 2009.

Jörg, B., Ojars K., Jeffery, K., & van Grootel, G. (2009b) CERIF XML 2008 – 1.0 Data Exchange Format. *euroCRIS*, April 2009.

Jörg, B., Jeffery, K., Asserson, A., van Grootel, G., Rasmussen, H., Price, A., Vestam, T., Elbæk, H. N., Voigt, R., Simons, E.J. (2009). CERIF 2008 – 1.0 Semantics. *euroCRIS,* 2009.

Storey, V.C. (1993) Understanding Semantic Relationships. *The International Journal on Very Large Databases (VLDB).* Volume 2, Number 4, October 1993, pages 458-488, Springer Berlin-Heidelberg.

Wand, Y. & Weber, R. (2002) Research Commentary: Information Systems and Conceptual Modeling—A Research Agenda. *Information Systems Research Journal*, Vol. 13, No. 4, December 2002, pp. 363-376.

Wang, R.Y., Storey, V.C., & Weber R. (1999) An ontological analysis of the relationship construct in conceptual modeling. *ACM Transactions on Database Systems (TODS) Journal*, Vol. 24, Issue 4, December 1999, pp. 494-528. New York USA.

UNISIST (1971) Study *Report on the Feasibility of a World Science Information System.* 171 pages, UNIPUB INc., P.O. Box 433, New York, N.Y.