

CERIF: Common European Research Information Format – Insight into the CERIF 2008 Release –

Brigitte Jörg

German Research Center for Artificial Intelligence, Saarbrücken

Summary

CERIF, the Common European Research Information Format, allows for a data-centric view on research information. It has developed from a data model that originates back to the eighties, and since then, major updates and improvements have been implemented. With the 2006 release, a substantial revision of the model towards the coverage of semantics with relationships has been incorporated and an XML-based interchange format has been introduced. The 2008 release elaborates on the Publication entity and opens the much requested connectivity to repositories of scholarly publications. This tutorial aims to give insight into the latest version of the CERIF model and format. It will demonstrate its power and flexibility for setting up quality research information systems and for enabling the quality interchange and integration of data at a large scale, and thus support the quality access to research information and services.

1 Introduction and Background

Most European countries collect and store their research information in national repositories. This research information is often spread over several regional or local systems with heterogeneous structure. In order to get additional value out of heterogeneous information sources, their inherent structures have to be mapped into a specified format that allows for a quality integration of the individual sources within a target system. Obviously, the integration of information at national level is not an easy task, and even more difficult at the European scale. Access to current research information and the exchange of research information at European level however is an essential requirement in the European Research Area¹ and for innovators, academics, decision makers, media, and the members of the society in general. Without a backbone format for research information exchange and storage it will be very difficult to provide quality access and added value services at a European range.

CRIS and CERIF approaches are not new. In the 1970s serious efforts for international cooperation among research information systems were being made to survey a country's scientific and technological potential, and to use such information in the formulation of science policy on a national level². In 1971, UNISIST³ published a "Study report on the feasibility of a world science information system" [UNISIST 1971].

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

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

³ UNISIST: UNESCO's World Scientific Information Programme

The first release of CERIF has been published in 1991 with the aim of facilitating data exchange of records on research projects between European Member States, and to serve as a format to allow for the networking of databases. The European Working Group on Research Databases has recommended the CERIF format as a result of a workshop held in 1987. The structure of CERIF 1991 was based on a data model describing “Research Projects” only. The needs for an extension were recognised. In 1997 revision work was entrusted to unit D2 DG XIII of the European Commission. The revision in the model was based on a reflection of user requirements and led to a recommendation for CERIF 2000 to Member States and a handover of CERIF to euroCRIS⁴. The CERIF 2000 release added Person and Organisation as entities and many other entities relevant in the research context, such as Publication, Service, Equipment, Patent, Country, Language, Event, etc., and Classification. Additionally, the entities had assigned types and their relationships allowed for roles to capture the semantics. With the CERIF 2006 release the role and type entities have been moved to the so called Semantic Layer, a generic solution across the model that allows for flexibility in capturing the semantics for different applications and views. With CERIF 2006 the Publication entity has become a core entity, and CERIFXML 2006 has been published as an interchange format. The CERIF 2008 release extends its predecessor with elaboration on the Publication entity for bibliometric studies and for a better connectivity with repositories for scholarly publications.

The CERIF tutorial will give insight into the CERIF 2008 model (entities and structure) with a major focus on relationships and the Semantic Layer, and the upgrade of the Publication entity. Additionally, the CERIF XML interchange format will be presented and discussed in detail.

2 CERIF 2008 Release

With the CERIF model upgrade in the previous release [Jörg et. al. 2007], we achieved a stable backbone structure and flexibility with capturing the semantics of relationships. According to their characteristic features, the CERIF entities are divided into five groups:

- (1) Core CERIF Entities
- (2) 2nd Level CERIF Entities
- (3) CERIF Link Entities
- (4) Language-dependent CERIF Entities
- (5) CERIF Classification Entities (Semantic Layer)

2.1 Core CERIF Entities

The core CERIF entities are Project, Person, OrganisationUnit, and ResultPublication. For each of them a screenshot from the Toad Data Modeler is provided that shows the entities and their corresponding relationships (Figure 1 – 4).

⁴ euroCRIS: <http://www.eurocris.org/>

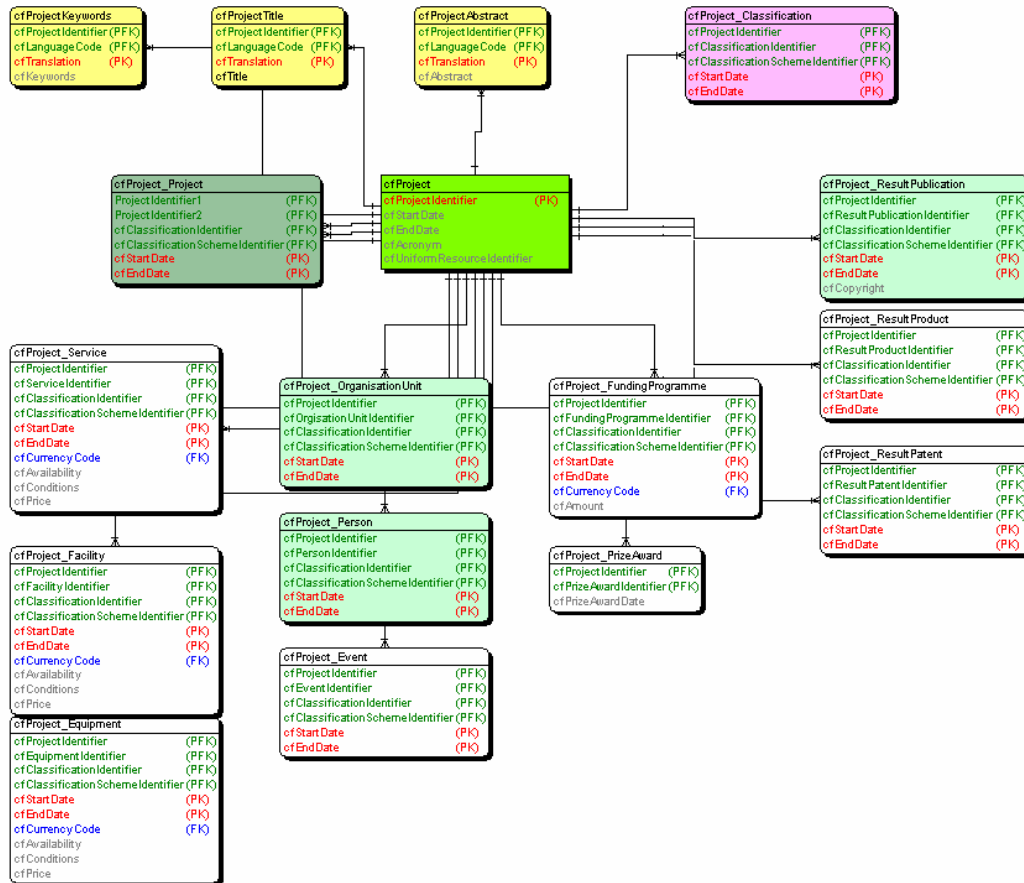


Figure 1: Full representation of the core entity Project and assigned relationships.

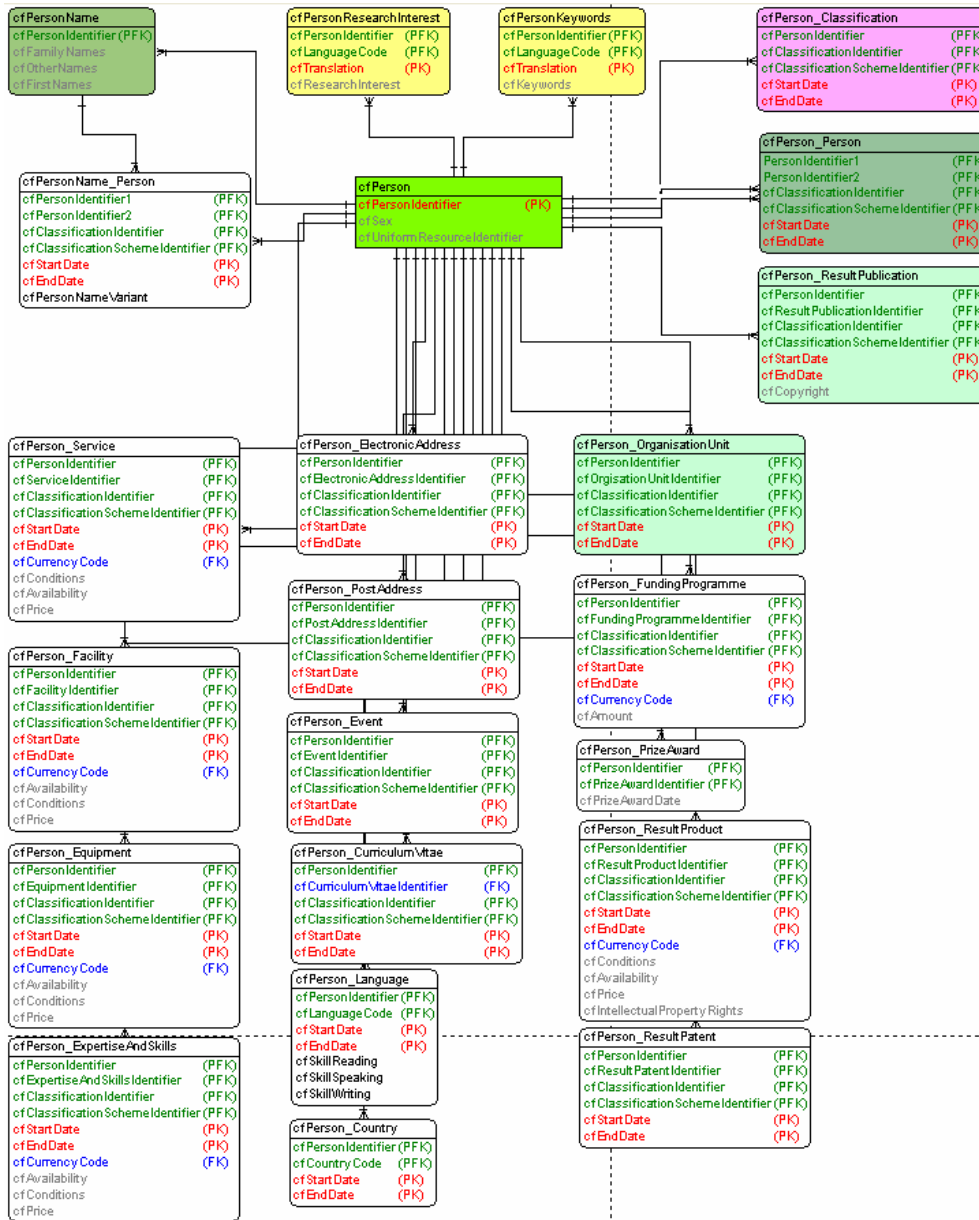


Figure 2: Full representation of the core entity Person and assigned relationships.

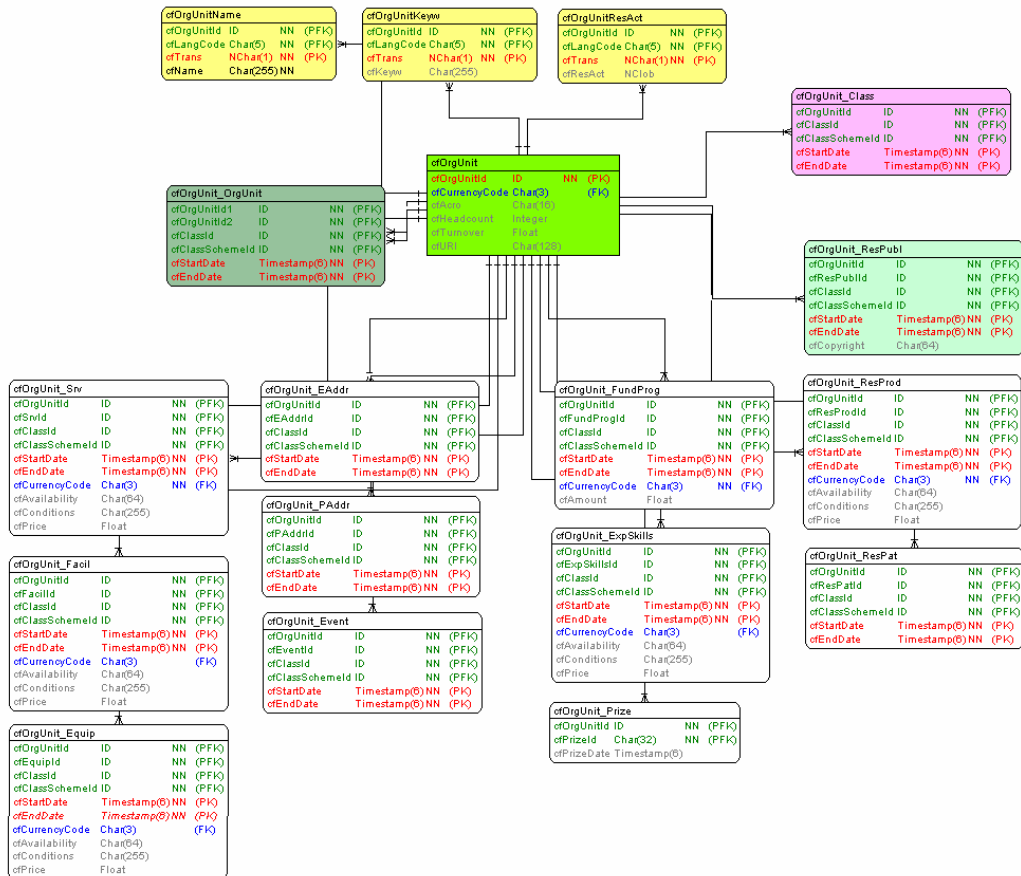


Figure 3: Representation of the core entity Organisation and assigned relationships.

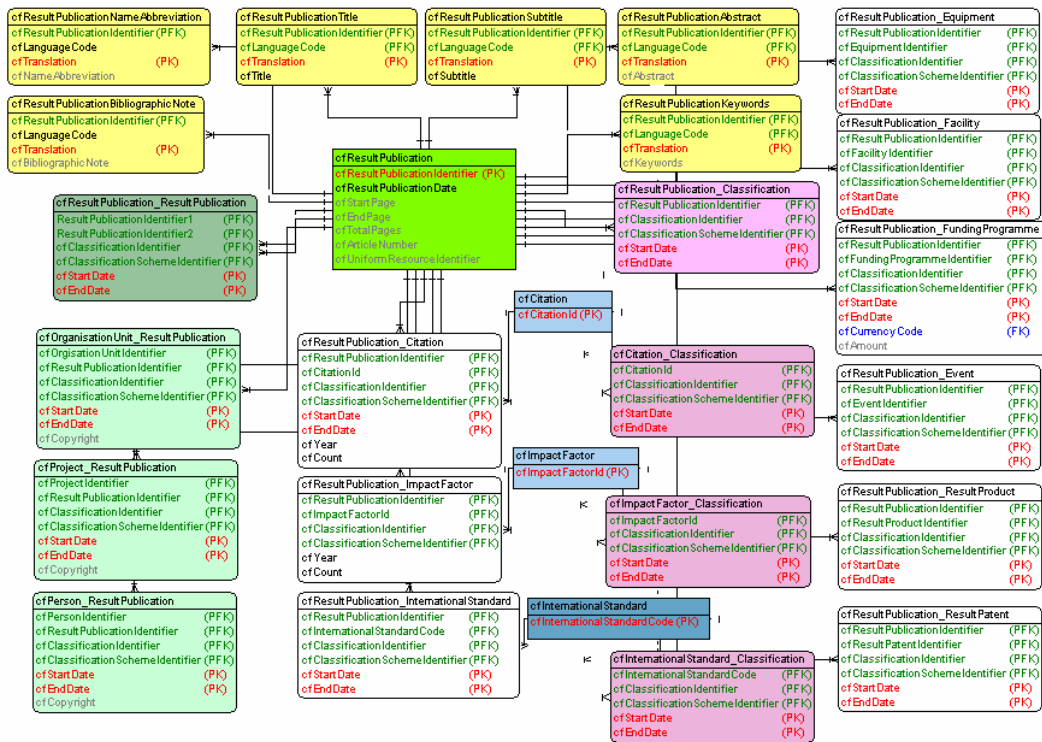


Figure 4: Representation of the core entity ResultPublication and assigned relationships.

The extension of the ResultPublication entity has been considered useful for quite some time and again recently with ongoing implementations of CERIF-based CRISs for scientometric analysis.

2.2 2nd Level CERIF Entities

CERIF 2nd level entities capture the context of activity and interaction in the wider range. Figure 5 shows the core entities and some of the 2nd level entities.

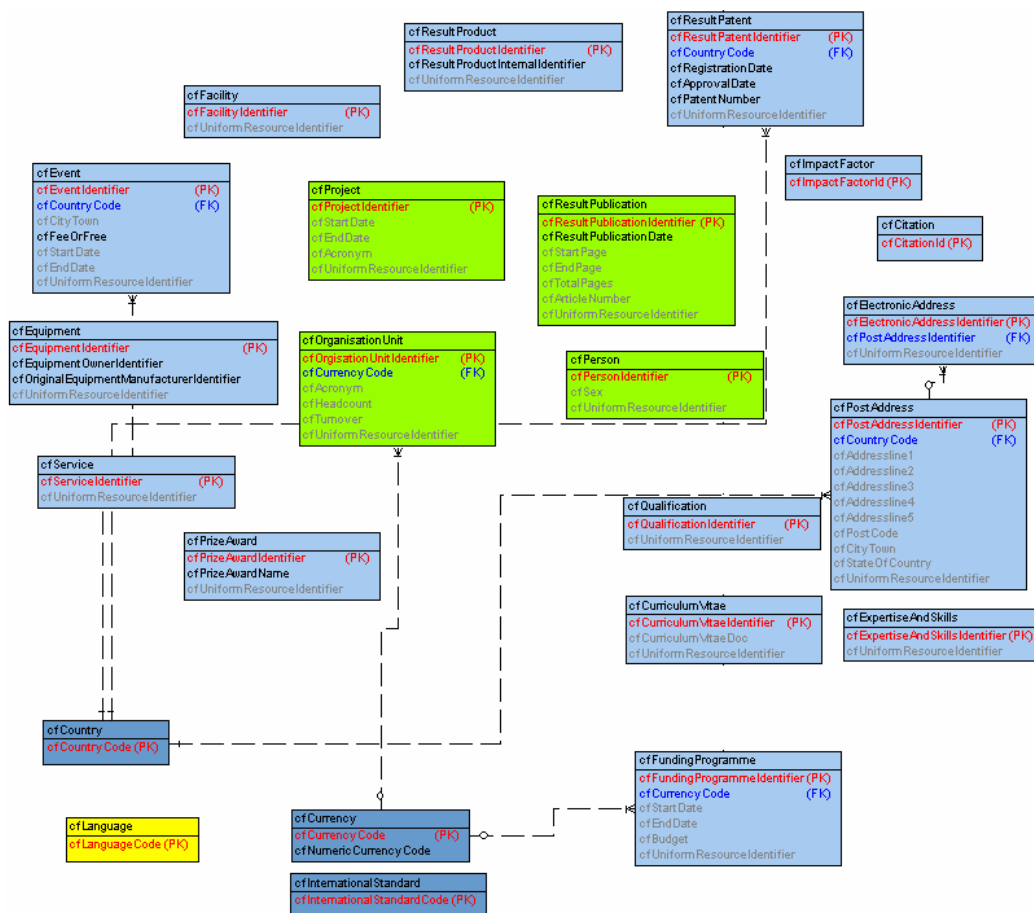


Figure 5: Representation of the four core entities and 2nd level entities.

2.3 CERIF Relationships (CERIF Link Entities)

The CERIF entities are linked with each other as relationship entities by their IDs. References to classifiers and particular schemes provide the flexibility with representations of the semantics. Timestamps allow for representations in any such composition that differs by date, as indicated in figure 6 and with previous representations of entities.

cfEntity1Name_Entity2Name
cfInheritedEntity1Identifier
cfInheritedEntity2Identifier
cfInheritedClassificationIdentifier
cfInheritedClassificationSchemeIdentifier
cfStartDate
cfEndDate

Figure 6: Structure of CERIF Link Entities.

2.4 Language-dependent CERIF Entities

Much information in research environments needs representation in more than one language. As indicated in figure 7, CERIF contains many language-dependent attributes, such as Keywords, Abstract, ResearchInterest, Name, ResearchActivity, Title and more. The semantic layer also allows for multiple language representations.

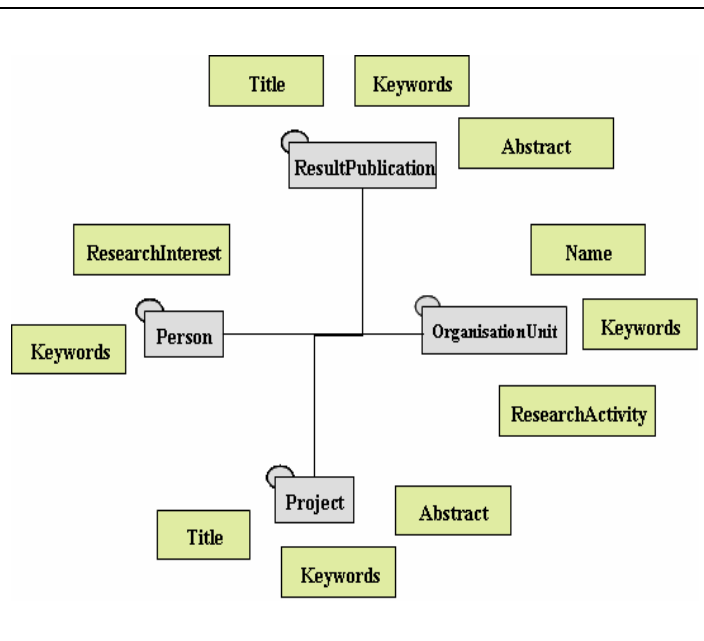


Figure 7: Language-dependent CERIF Entities.

2.6 CERIF Interchange Format (CERIFXML)

The CERIF XML interchange format allows for data exchange operations at physical level, and therefore conforms to the physical structure of the CERIF model (compare figure 9 and 10).

```
<cfOrgUnit>
  <cfOrgUnitId>ID</cfOrgUnitId>
  <cfURI>String</cfURI>
  <cfAcro>String</cfAcro>
  <cfHeadcount>Integer</cfHeadcount>
  <cfTurnover
    currencyCode="EUR">Float</cfTurnover>
</cfOrgUnit>
```

Figure 9: OrgUnit XML representation

cfOrgUnit	
cfOrgUnitId	(PK)
cfCurrencyCode	(FK)
cfAcro	
cfHeadcount	
cfTurnover	
cfURI	

Figure 10: OrgUnit relational entity representation

3 Next Steps

The CERIF 2008 model as presented will be released after members' review in June/July 2008. Explicit semantic values for the core entities' relationships will be presented during the tutorial at the conference in Maribor.

4 References

- 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.
- Aksnes, Dag W.; Revheim, J.-B. (2000): *The application of CRIS for analysing research output – problems and prospects*. In Proceedings 5th International Conference on Research Information Systems 2000. Kassel, Germany.
- Erbach, G. (2006): *Data-centric view in e-Science information systems*. In: Data Science Journal, Vol. 5, 2006, pp 219-222. Online ISSN: 1683-1470.
- Jörg, B.; Krast, O.; Jeffery, K.G.; Van Grootel, G.; (2007): *CERIF 2006XML – 1.1 Data Exchange Format Specification*. euroCRIS, April 2007
- Jörg, B.; Jeffery K.G.; Asserson, A.; Van Grootel, G.; Grabczewski, E. (2007): *CERIF2006 1.1 Full Data Model (FDM) – Model Introduction and Specification*. euroCRIS, October 2007.
- Jeffery, K.G.; Asserson, A.; Lopatenko, A. S. (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.

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.

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.

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, pages 494-528. New York USA.

W3C Recommendation: *Extensible Markup Language (XML) 1.0*, Fourth Edition, 16 August 2006, edited in place, 29 September 2006. <http://www.w3.org/TR/2006/Rec-xml-2006-08-16/>

W3C XML Schema: <http://www.w3.org/XML/Schema>

5 Contact Information

Jörg Brigitte

German Research Center for Artificial Intelligence

Language Technology Lab

Stuhlsatzenhausweg 3

66123 Saarbrücken

e-mail: brigitte.joerg@dfki.de

<http://www.dfki.de/~brigitte>