CERIF-2000 as a platform for university public research information service

A. Lopatenko, Vienna University of Technology

The CERIF-2000 set of recommendation specifies relational database schema and metadata schema which can be used to describe research information. In the project Austrian Research Information system – Multimedia Extended (AURIS-MM) the CERIF-2000 database was used as a database schema to keep information about universities' research. In the article a set of requirements to university information system is described with description how CERIF-2000 suits those requirements. The advantages and disadvantages of CERIF-2000 are emphasized and solutions to solve possible problems are suggested. The aim of the article is to provide very practical description of CERIF-2000 as a database for university CRISs.

Introduction

Wide public and professional interests to university research can not be satisfied by full-text search using internet search tools over researchers' and teams' web pages. It requires more sophisticated technologies to store and provide access to research information – such technologies as relation databases, and web interface for attributed search of structured information. In Austria most university already have own or bought information system for research information.

The full development of each system from database design to ready system requires a lot of efforts and often it is too costly. Providing access to research information at national level requires integration of information from different universities, what is often also too costly if systems were developed independently based possibly on different view.

In the development of AURIS-MM which stores information from a number of Austrian universities attempt to use already developed database and set of guidelines were done to make the system development cheap, maintainable, re-usable and easy to integrate with future European systems.

The CERIF-2000 was taken as a core database schema. A number of extensions to CERIF-2000 were developed to satisfy particular university CRIS requirements. A number of corrections of CERIF-2000 errors was done t make it error-free database. The error-free release of CERIF-2000 is available for public.

The AURIS-MM was developed as Java Server Pages application, and tested on Tomcat platform

Requirements to university public research information system

The AURIS-MM was built as a system to provide access to public research data of Austrian universities. The first main aim of the system is to provide unified search and navigation access to consolidated research information without update/change/input interfaces for individual researchers. The second aim of the system is to provide a number of services such as notification service, collection management service which assist information consumers in their main routines of information processing.

Analysis of university systems already built in Austria and other countries, as well analysis of publication of previous EuroCRIS conferences helped to recognize main requirements to public research CRISs

A. full-text search interface for all information in the database with support of sophisticated full-text query operators

B. attributed search interface for each type of entity
C. effective use of keywords and thesaurus in search operations
D. ability to have some common reasoning about information under condition of a lack of information
E. sophisticated navigation access for exploration of information
F. data schema must be compatible with old data and common view on research information

CERIF-2000 as a CRIS platform

AURIS-MM information system uses CERIF-2000 database schema as underlying core database schema extending it when it is needed. A number of errors (about 60) were found in original CERIF database schema. The errors were corrected and error-free version of CERIF database is published. The new version of CERIF-2000 database were checked for errors in Oracle designer, by running database creation scripts in Oracle and by using it as a database for AURIS-MM. The exploration of new schema showed that it suits requirements for database development.

The use of CERIF-2000 has shown that all previous requirements can be satisfied with minimal efforts

- full-text search interface for all information in the database with support of sophisticated full-text query operators

The CERIF-2000 database very clearly specifies in which fields/table columns full-text description should be stored. Use of special software like Oracle InterMedia makes possible to implement sophisticated full-text search services.

- attributed search interface for each type of entity

Due to high-normalized database structure of CERIF-2000 real research information search queries are expressed as database queries.

- effective use of keywords and thesaurus in search operations

Experience shows that high-professional users of research information intensively use keywords from subject specific vocabularies, thesaurus in search operations. CERIF-2000 separates subject classifications (vocabularies), as well “keywords” from full-text description for research projects. For organizations, persons CERIF allows to describe objects by classification terms, “expert skills” what make possible implementation of keyword-search service

- capability of common reasoning about information under condition of a lack of information

The development of database which stores data from multiple universities showed that due to different views on data, some data maybe very poor from global system view. For example, “collaborator search” services require detail description of research interests and experience organizations (part. institutes), what is not implemented in most Austrian university CRISs. But, for example, project descriptions of organization can be used to find out organization expert skills. AURIS-MM defined a set materialized views and views which utilize information about projects for persons and organization units.

- sophisticated navigation access for exploration of information

Highly-normalized structure of CERIF-2000 database, as detail descriptions of relations between objects (attributed m-n relation with roles, dates) allows to describe relation in details in use them for navigation services.

- data schema must be compatible with old data and common view on research information

The use of CERIF with Austrian university research information has shown a high compatibility of CERIF with established practices of research information description for needs of public and research administration. Really in public part of research information CERIF schema is a superset of any other schema (excluding description of publications).
CERIF disadvantages
Problem: CERIF is very high-normalized database schema. Some search operation requires a join of multiple tables (up to 7-8 in AURIS-MM) what makes them very slow
Solution: a number of materialized view or new tables which provides query oriented view on information stored in database make queries easer and faster.
Problem: to provide fast access for full-text search with sophisticated query operators to CERIF database physical database design is needed which is not a part of CERIF-2000
Solution: a number of indexes is proposed which makes possible easy construction of full-text queries. The indexes are proposed for Oracle indermedia search engine
Problem: a number of services needed for university CRIS are not supported by CERIF database schema
Solutions: extensions to CERIF database which maybe applied to CERIF schema are proposed
Problem: information in different countries and communities is described by different vocabularies, which maybe different from CERIF vocabularies
Solution: CERIF classification tables allow to store as many vocabularies as needed (with some restrictions on interterm relations - CERIF supports only broader/narrower term relations). To use local vocabularies, it suggested to store local vocabularies in tables. The mapping between local vocabulary terms and CERIF vocabularies must be described in special tables to make CRISs compatible with CERIF. To make CRIS still compatible with other CRISs based on CERIF during data exchange operations vocabulary mapping is used to produce metadata in CERIF format described by CERIF vocabularies

Conclusion

The use of CERIF-2000 as a core database schema for university CRIS shows that CERIF-2000 is useful and easily to maintain and extend. High normalization of database makes it highly compatible with other different approached to description of research information. Comparison of CERIF-2000 with other Austrian university databases, developed independently of CERIF showed that CERIF view on information is coincide with view of other developers and information consumers.
The CERIF-2000 has some disadvantages, most of which can be eliminated or corrected in system development. The development of university CRIS using already published experience and error-free database schema can be done in a few man-months