

# **CERIF - Information Retrieval of Research Information in a Distributed Heterogeneous Environment**

Andrei Lopatenko UM, Anne Asserson, UiB, Keith G Jeffery CLRC

## **Summary**

User demands to have access to complete and actual information about research may require integration of data from different CRISs. CRISs are rarely homogenous systems and problems of CRISs integration must be addressed from technological point of view. Implementation of CRIS providing access to heterogeneous data distributed among a number of CRISs is described. A few technologies – distributed databases, web services, semantic web are used for distributed CRIS to address different user requirements. Distributed databases serve to implement very efficient integration of homogenous systems, web services - to provide open access to research information, semantic web – to solve problems of integration semantically and structurally heterogeneous data sources and provide intelligent data retrieval interfaces. The problems of data completeness in distributed systems are addressed and CRIS-adequate solution for data completeness is suggested.

## **1. INTRODUCTION**

One of the challenges of CRIS development is to provide access to research data which are scattered on the web pages, stored in different research information systems. The informational needs of a researcher or policy-maker are very seldom limited to information from one CRIS, which usually represents research from a particular region or sector of science. There is a strong need to integrate information from different sources and to provide access to all information to users, enabling them to utilize a wide range of sources. One of the problems of such system development is heterogeneity of research information systems. The CERIF Task Group develops the CERIF system to solve problems of providing transparent access to disparate research sources

The paper is organized as follows. Section 2 outlines requirement for research data integration and describes results already achieved in this area. Section 3 describes which approaches are being developed by CERIF Task Group. Section 4 described user demands for quality of data and suggests solutions for distributed data queries to achieve data completeness. Section 5 concludes

## **2. REQUIREMENTS FOR RESEARCH DATA INTEGRATION**

Typical needs of research information consumers are not limited to information from one research information system. The information about research in the same area, about different stages of one research, about different research relevant information are scattered among different information systems. Getting the information requires knowledge of where which information is stored and efforts to visit all systems, learn them and use them to find required information. Of course, when the systems are not integrated, information from one system cannot be reused in another or by the user system automatically. Furthermore, the information consumer must know all differences in description of research information, vocabularies used - which makes retrieving of information often an intractable task.

Full-text search engines, like Google, AskJeeves, widely used to search information on the web:

- do not go inside many information systems
- do not provide attributed search which is important for policy-makers and researchers search
- do not filter information by its quality, actuality
- do not provide facilities to use terms, thesauri specific for science and which are very important for research informational retrieval
- do not manage the semantics of query and information searched

The strong requirement to have the ability to search research information from all European sources led to the development of the prototype project ERGO – European Research Gateways Online. This project, organized and sponsored by European Commission, aimed to collect metadata information about research projects from all European countries into one central database and to provide information retrieval access to that database to users. The ERGO project, in which the euroCRIS group strongly participated, created a data exchange approach to collecting information based on SGML technologies.

The need to integrate distributed research data and provide unified access was recognized in the Santa Fe Convention (VanDeSompel, 1999). This led to the development of the Open Archives Initiative – an approach which consists of a metadata set and protocol to integrate distributed metadata on the web, and which is already used by several communities. The need to provide access to data from distributed information systems of scientific libraries led to the creation a number of distributed systems such as MathNet (MathNet) PrePrint (PrePrints) network, Networked Digital Library of Thesis and Dissertation mentioned(NDLTD, NDLTD-01, NDLTD-01), Research Papers in Economics (RePEC), Clinical Medicine NetPrints (NetPrints) . A number of other examples show a strong need for distribution information retrieval in science. The requirement for informational retrieval from multiple data sources demands resolution of semantic, structural and system heterogeneities.

A lot of research and development were done in development of distributed systems for information retrieval but not all maybe applied to CRISs.

Requirements to distributed solution for CRIS:

1. **Easiness to implement.** The technology used should be well-known, the solution must be implemented with minimum demand for financial, human and time resources. Due to this reasons it was decided to implement free toolkit, which maybe installed and configured. The toolkit is implemented on Java and well-documented.
2. **Flexibility.** CRISs usually are not static systems, they are very changeable. The system should accommodate changes. The solution should be easily configured for new CRISs. A few solutions for distributed CRIS are suggested. Each solutions is address different types of requirements for distributed CRIS and different type of data sources. We tried to develop software architecture making configuration of the system easy.
3. **Support of open standards.** Very probably many CRISs will be embedded into enterprise or university infrastructures, research information should be possible to integrate into established data flows. Also reuse of team experience and software tools makes CRIS development and maintains cheaper. Now some of the most popular standards or activities

related to data access in software development are JDBC, XML, RDF, SOAP, Java Beans (+EJB), CORBA. Current version of distributed CERIF may use JDBC, XML, RDF, SOAP.

4. **Effectiveness.** Seeking of research information is data and reasoning intensive operations. Users expect getting data on their request very fast. A solution for fast data search based on distributed database technologies is suggested.
5. Ability to solve problems of **semantic and structural interoperability**. The CRISs are very heterogeneous. Despite their heterogeneity uniform access to data should be provided. Semantic Web solution for CRIS address a wide range of semantic diversity problems. The meaning of data may be included into the systems and used for search

### 3. CERIF DISTRIBUTED INFORMATIONAL RETRIEVAL SOLUTION

The CERIF Task Group decided to develop different solutions, because no one solution can be enough to solve all problems of transparent access to disparate data sources. Different requirements of a distributed solution, and the technology used, necessitate different approaches. Different approaches also are diverse in efficiency, compatibility with legacy technologies, ability to solve problems and so real systems might require use of all of them.

The new solutions being explored and developed are in addition to existing solutions utilising CERIF such as:

- simple portal pointing to CRISs (e.g. DRIS);
- central catalogue (possibly replicated-distributed) and contact information for individual CRISs e.g. ERGO Pilot
- central catalogue and automated retrieval from CRISs (ERGO 2++ proposal);
- central catalogue and advanced knowledge-assisted retrieval from CRISs (ERGO3 proposal)

To nurture the CRIS community, help CRIS developers and finally give better services to researchers and others, the principal policy of the CERIF Task Group is to provide free source code for solutions and develop them based on open standards and technologies. The source codes and documentation with demonstration are available at (CERIF TG)

#### 1.1 Distributed Database Approach (CERIF-DD)

The aim of this approach is to provide very efficient search of multiple databases, when semantic and structural heterogeneity is minimal, the systems use common thesauri, or thesauri mappings are provided. This approach does not solve automatically semantic or structural heterogeneity problems, but is characterized by efficiency and is easy to develop and deploy. The most likely application of this approach is unification of research databases belonging to different branches of one organization. Each database must export a set of views conforming to the requirements of CERIF (signature of view and semantics of attributes). JDBC access to database views for select operations must be provided. For each network one central server repository exists which registers all databases and provides a search over all databases to the end user. When views are exported for a database, the database must be registered at the central server of network, with description of source of data, quality, actuality,

sector of research and access information. After the database is registered it will be used by the central search facility to answer queries of users. The data source should be registered at the central server through SOAP web service interface just sending RDF description of data source.

The set of view definition for CERIF-DD is precisely defined (CERIF 4). The signature of view and semantics can be clearly understood from CERIF-DD view definitions and they are described at CERIF TG pages. Currently the search for mostly used entities like persons, projects, organization units, publications and links between them maybe performed in CERIF-DD

Data source should be described in RDF according CERIF source description DAML Schema (CERIF 3)

This solution makes possible to solve manually very simple semantic and structural heterogeneity problems, just adequate definition of views for data source must be provided which maps local tables and columns into right views and view attributes. Investigation of some custom CERIF independent university CRIS has shown that such mapping is possible in most cases what says about high CERIF compatibility with CRIS (Lopatenko 2001.1, Lopatenko 2001.2)

Also this solution makes possible to solve problems of usage of different vocabularies in different CRISs. When the local data source use different from CRISs network vocabularies for describing data only table definition of vocabulary mapping should be provided and this mapping should be included into view definition. So if such mapping exists, then search operation will find adequate data even if they described by different vocabulary

Another advantage of this solutions is due to relational access to data and use of JDBC tools for data analysis, warehousing, decision making based on this technologies could be used

Disadvantages: possible security problems with direct access to database (must be guarded), high demand for manual configuration (view definitions, vocabulary mapping look-up tables), solving only simple semantic interoperability problems, incompatibility with web infrastructure

## **1.2 Semantic Web Approach (CERIF-SW)**

The Semantic Web (SemWeb, TBL 2001) approach aims to solve problems of distributed information retrieval when:

- structure and semantic are different in different data sources and search must take into account differences of data sources
- sophisticated informational retrieval operations are demanded
- the schema of existing and new data sources can be changed
- already published Semantic Web data must be used
- direct access to database cannot be provided
- compatibility with other SW based architectures is important

The Semantic Web approach is based on CERIF as the core ontology (CERIF 1) to describe the meaning of research data. CERIF is the base vocabulary which provides a common set of terms and which must be understood by any system conforming to CERIF-SW approach.

One of main features of Semantic Web solutions is ability to easy integrate very heterogonous data sources – digital libraries, databases, repositories, legacy sources. To add new CERIF based CRIS to network it would be enough to install CERIF-SW application, configure it to get data from local database and publish on the web.

When systems with different structure and meaning of data from CERIF must be integrated into the network, the ontology of systems must be described in CERIF terms if that is possible. It allows integration also over search of not-CERIF but CERIF-compatible data sources (DublinCore, MathNet, etc). Each data source must publish its ontology on the web and express data in RDF according to its ontology. The ontology must conform DAML(DAML) format and to make data searchable by CERIF queries, must explain local terms in CERIF terms like, for example, IST project is a Project, IST project is part of EU programme, financed by CORDIS, done in Information Technologies

To make data accessible for queries they must be transformed into RDF (use of CERIF toolkit is possible). The RDF data may be published on the web to directly access through http (for harvesting), may be accessed through web services (see next, very integrated with CERIF-WS approach), or may be published on RDF Networked Query Facility – a Network Query Engine for RDF data developed for CERIF, based on HP Jena toolkit. Then data source should register its data and ontology at search servers, which provide transparent access to data to end users.

Another one of main features of Semantic Web solutions of CERIF is ability to use ontology driven **intelligent information retrieval**.

The same information can be represented in information system in different ways, described from different point of view due to diversity of data sources, policy-restrictions, or even the position of the author / compiler. The views of the same information can be very different for different categories of information consumers. The ontology-driven information retrieval is intended to

1. solve problems of discrepancies between the viewpoints of information consumers and description of information in the system;
2. provide more powerful and sophisticated retrieval facilities, allowing to information consumers to utilize domain knowledge; make it possible for information consumers to investigate in which terms information is described in the system, what is the meaning of
3. those terms and if it is needed using this knowledge to create new more focused queries

### **Example**

The following example shows how ontological descriptions of the terms of an information system can be used to find out relations between those terms to create queries which satisfies user needs (completeness and relevance of returned information)

This ontology is an example ontology, not real one, some facts maybe wrong from political point of view or some types of ontology design. It is just a demonstration.

The demonstration is a database of European projects, programme, organizations. It stores information about objects - such as programmes, projects - each objects is described as belonging to some of the classes.

Base ontology

There are two types of classes - defined and primitive.

If it is said that the FundingOrganization is an organization, which finances some activities and it is also said that class FundingOrganization is primitive class then it means each FundingOrganization finances some activities, but if we know that X is an organization and X finances activities it is not enough to say that X is a FundingOrganization

But if it was said that FundingOrganization is a defined class then each object which is known as organization and which also is known to be financing some activities, it would become FundingOrganization in information system

Table 1. Ontology of application for searching projects

Classes	Definition
Activity	
Programme	
EUProgramme	Programme which is <u>financed-by</u> EU or by any EU-country
ISTProgramme	one of EU Programmes
Location	Any location can be a <u>part-of</u> another location
City	
Continent	
Europe	
Country	
EuropeanCountry	A country which is a <u>part-of</u> continent Europe
EUCountry	A country which is a <u>part-of</u> EU
Union	Geopolitical or economical union of countries
EU	One of unions, but which is a <u>part</u> of Europe
Organization	
FundingOrganization	This organization which finances some Activities or Projects
EUFundingOrganization	Any funding organization which is <u>situated-in</u> EUCountry
EuropeanFundingOrganization	Funding organization which is <u>situated-in</u> Europe
Project	
FinancedProject	Project which is <u>financed</u> by some of FundingOrganization
EUFinancedProject	Project which is <u>financed-by</u> EUFundingOrganization
EuropeanFinancedProject	Project which is <u>financed-by</u> EuropeanFundingOrganization
ISTProject	Project which is a <u>part-of</u> IST Programme
Properties	
financed-by	A <u>financed-by</u> Y reverse relation finances When A is <u>financed-by</u> Y then Y finances A
Finances	
part-of	transitive relation. when X is a part of Y, and Y is a part of Z, then X is a part of Z
situated-in	Geographical inclusion. transitive relation
Axioms	
Project which is a <u>part-of</u> any of EUProgrammes is <u>financed-by</u> EU	When it is known that a project is a part of EUProgramme then we are sure that this project is financed by one of EUFundingOrganizations

After ontology verification and classifying its terms, the verifier asserted new statements about relations between classes. Some of them:

Table 2. Automatically inferred conclusions from ontology

Statement	Proof
-----------	-------

ISTProject is an EUProject	<ol style="list-style-type: none"> <li>1. fact: ISTProject is a part of ISTProgramme,</li> <li>2. fact: ISTProgramme is a EUProgramme</li> <li>3. statement: ISTProject is a part of EUProgramme</li> <li>4. 3 + Axiom -&gt; ISTProject is financed by EU</li> <li>5. 4 + definition of EUFinancedProject -&gt; IST is an EUFinancedProject</li> </ol>
EUFinancedProject is an EuropeanFinancedProject	<ol style="list-style-type: none"> <li>1. EUFinancedProject is financed by FundingOrganization which situated in EU</li> <li>2. EU is situated in Europe</li> <li>3. 1 + 2 + transitivity of situated-in EUFinancedProject is financed by organization which is situated in Europe</li> <li>4. 3 + definition of EuropeanFinancedProject -&gt; EUFinancedProject is an EuropeanFinancedProject</li> </ol>

Such reasoning performed by Description Logic can

1. increase knowledge of the data of users of the information system, helping them understand more precisely the data stored in the system and their own needs (user gets information that IST project is EU project, despite it was not directly asserted in the system - this can help a user in search of IST projects )
2. provide mechanism for implementing more correct query facility which utilizes new knowledge to find more complete answers on users' requests - eg. on request of IST projects, the databases of EU projects will also be searched
3. provide facility to describe results of answer to user - eg. why IST project was returned on EU projects request

As reasoning facility CERIF-SW uses FaCT system (FaCT) – Description logic classifier developed in Manchester University. As user friendly tool to develop CERIF ontology and other ontologies for CERIF solutions, ontology editor OilED(OilEd), developed in Manchester University was used

So, main advantages of Semantic Web solution for CERIF

- ability automatically deal with heterogeneous data sources
- ontology-driven intelligent information retrieval
- compatibility with a set of emerging projects, activities
- ability to implement knowledge management solutions
- no security problems of direct access to database
- maybe build over any, not only database system allowing to utilize legacy systems, knowledge management systems, workflow support systems and others

Main disadvantages of CERIF-SW

- very inefficient now, search operation on database over tens megabytes and requiring transfer of a lot of data may not satisfy fast access requirement
- effective use of technology requires from developers knowledge of Semantic Web, ontologies and some tools – additional resources needed to implement
- warehouse, analysis and other reporting tools developed for relational databases can not use data from SW directly.

### 1.3 Web Services Approach (CERIF-WS)

The Web Services (W3 Web Services, IBM Web Services, Glass-2000) approach aims to solve problems of distributed informational retrieval

- important to make research information sources compatible with emerging standards of corporate networking and put research data into enterprise information flows
- demands to provide Web Services to research data, which are becoming popular for corporate internetworking
- direct database access is impossible for security or other reasons
- XML/SGML is already used in the organization and any new solution must be compatible with those already developed. Also experience of the team is a great asset and new solution must be based on such experience
- to provide efficient transport level protocol for CERIF-SW
- searched CRIS are CERIF-compatible (at intensional level) but have very different database schemas

CERIF-WS consists of

- an XML Schema based on the CERIF ontology;
- a SOAP (Simple Object Access Protocol) implementation of wrappers to each data source which implements functions to search and retrieve data;
- the CERIF-WS search facility, which searches registered SOAP services and provides access to research data from all systems to the user;

CRISs with a schema different from CERIF can easily participate in the CERIF network by publishing (through a wrapper) their data in CERIF XML. CERIF-WS has been built in a such way that it is very easy for developers to create or change the XML encoding of their data in the wrapper. Thus the technique allows CRISs with non-CERIF encoding to be accessed as if they were CERIF-encoded at the expense of building and maintaining the wrapper.

The main advantages of Web Services are

1. implementation of WS layer for CRIS maybe very easy. The tools to support WS for a lot of languages, platform are already available.
2. it is possible that in near future Web Services will dominate as a standard for interoperability of systems in business applications (Gartner 2002) – research information maybe access by other business applications
3. Web Services solution for CERIF may utilize Semantic Web for sophisticated queries, integration of heterogeneous data sources

Disadvantages of Web Services for CERIF

1. now solutions based on them are not very efficient comparing to distributed data search
2. resources to teach team use Web Services are needed
3. Web Services are not yet very mature, some tools are not efficient or have a lot bugs
4. current CERIF implementation of Web Services allows to use only very simple queries, hard to implement sophisticated queries which requires to investigate relations between objects

#### **4. Data Quality**

It is expected to have high demand for quality of scientific data in some CRISs. We believe that main parameters of quality of scientific data are completeness, actuality and correctness. Completeness of data in CRIS is presence of data about all entities which are subject of given CRIS. For example, CRIS about European projects since 1991 in chemical research in Norway will be complete, if it contains data about all European projects in chemical research since 1991 in Norway

Actuality of data in CRIS is presence of newest information about CRIS subject. For example, CRIS about research projects in university will store actual data, if data in CRIS will be updated on new project developments immediately.

Usually correctness, actuality and completeness of data in given CRIS are results of administrative efforts for data harvesting, analysis, input.

But in distributed case, when data on request are returned from a few different independent information system, for quality-critical applications reasoning procedures to judge quality of data are required

A few approaches to judge about quality of answers in distributed systems or to get complete answers for some queries are investigated (Levy 96, Motro 89, Duschka 97, Enzioni 94, Minok 99). It was decided to accommodate in CERIF-2000 distributed architecture (Duschka 96) because it satisfies information needs of CRIS systems, well-described and feasible to implement. (Minok 99), for example, approach requires existence of universal relation what is not a case of CERIF-based CRIS.

To make possible reasoning if query answers are complete metadata about completeness (local completeness) of sources must be provided according to CERIF registry metadata schema (CERIF 3) The actuality and correctness of data are not yet addressed in distributed CERIF project

#### **5. CONCLUSION**

CERIF has demonstrated the basic soundness of the datamodel both in formal correctness and in its designed-in flexibility, as well as compatibility with developed and emerging CRISs. The authors' investigation of formats for scientific data shows high compatibility of CERIF with new systems and networks especially in the various W3C (World Wide Web Consortium) development groups. Extensions and solutions being built for CERIF by the CERIF Task Group provide implementations of importance for the research community information services, like distributed information retrieval and semantic informational retrieval.

#### **References**

CERIF TG <http://www.eurocris.org/cerif>

CERIF 1 DAML ontology for research information <http://www.eurocris.org/cerif.daml>

CERIF 2 XML Schema for research information <http://www.eurocris.org/cerif/cerif.xsd>

CERIF 3 DAML Schema for data source metadata <http://www.eurocris.org/cerif-sources.daml>

CERIF 4 Set of view definitions for distributed database solution <http://www.eurocris.org/cerif/cerif-dd.sql>  
DAML <http://www.daml.org>  
FaCT <http://www.cs.man.ac.uk/~horrocks/FaCT>  
Gartner 2002 Gartner Says Web Services Will Dominate Deployment of New Application Solutions for Fortune 2000 Companies by 2004-Reports Examine The Future of Web Services And Vendors Driving The Industry, Jan. 2002, <http://industry.java.sun.com/javaneews/stories/story2/0.1072.41782.00.html>  
Duschka 97 Duschka O., Query Optimization Using Local Completeness, In Proc. of 14<sup>th</sup> AAAI National Conference on Artificial Intelligence, AAAI-97., July 1997  
Enzioni 94 Oren Enzioni, K. Golden, D. Weld, Tractable closed world reasoning with updates, In. Proc. 4 Knowledge Representation  
Glass-2000 Glass G.; The Web services (r)evolution, Part 1, <http://www-106.ibm.com/developerworks/webservices/library/ws-peer1.html>  
IBM Web Services <http://www-106.ibm.com/developerworks/webservices/>  
Levy 96 Levy A., Obtaining Complete Answers from Incomplete Databases Proceedings of the 22nd VLDB Conference, Bombay, India. 1996  
Lopatenko 2001.1 Comparison of CERIF-2000 and Salzburg University research information system schemas, <http://derpi.tuwien.ac.at/~andrei/documents/SzbCERIF.htm>  
Lopatenko 2001.2 Comparison of CERIF-2000 and AURIS (AUSTrian Research Information System) schemas, <http://derpi.tuwien.ac.at/~andrei/documents/AURIS-CERIF.htm>  
MathNet <http://www.math-net.de>  
Minok 99 Minock M; Rusinkiewicz M.; Perry B; The Identification of Missing Information Resources by using the Query Difference Operator, Proc. of the 4th International Conference on Cooperative Information Systems  
Motro 89 Motro M.; Integrity = Validity + Completeness. ACM Transaction on Database Systems, TODS 14(4): 480-502 (1989)  
NetPrints <http://clinmed.netprints.org/>  
NDLTD <http://www.ndltd.org/>  
NDLTD-01 Suleman H.; Atkins A.; Goncalves M., France R.; Fox E.; Chachra V.; Crowder M.; Young J. Networked Digital Library of Theses and Dissertations, Bridging the Gaps for Global Access - Part 1: Mission and Progress, D-Lib Magazine, Vol. &, Num. 9, Sep. 2001  
NDLTD-01.1 Suleman H.; Atkins A.; Goncalves M., France R.; Fox E.; Chachra V.; Crowder M.; Young J. Networked Digital Library of Theses and Dissertations, Bridging the Gaps for Global Access - Part 2: Services and Research, D-Lib Magazine, Vol. &, Num. 9, Sep. 2001  
OilEd <http://oiled.man.ac.uk/>  
Preprints <http://mathnet.preprints.org>  
RePEc <http://www.repec.org/>  
SemWeb <http://www.w3.org/2001/sw/>  
TBL 2001 Berners-Lee T., Hendler J.; Connolly D.; Swick R.; The Semantic Web, Scientific American, May 2001  
Van de Sompel H.; Lagoze C.; The Santa Fe Convention of the Open Archives Initiative, D-Lib Magazine., Feb 2000, Vol. 6., Num. 2  
W3 Web Services <http://www.w3.org/2002/ws>