

CRIS: Research Organisation View of the e-Infrastructure

Keith G. Jeffery, Anne Asserson

STFC Rutherford Appleton Laboratory Didcot, University of Bergen

Abstract

The e-infrastructure provides interconnection of computing and data resources. Low-level middleware has been and is being developed to conceal the heterogeneity of the physical e-infrastructure and thus provide a platform for information services. The information services form the i-infrastructure, supported by the underlying e-infrastructure.

A CERIF-CRIS (Common European Research Information Format – Current Research Information System) provides an end-user with access to information about the research at their organisation. Commonly the CERIF-CRIS is linked with a repository of scholarly publications holding the full text (or multimedia) and with a repository of research datasets and software. In both cases the CERIF-CRIS provides both metadata and context. The CERIF-CRIS may also link to other systems at the organisation such as finance, human resources, project management and thus provide a view into those systems relevant to the needs of CRIS users. The CERIF-CRIS may be used to provide authentication and authorization of users, to control workflow of e-processes and to provide directory services. The CERIF-CRIS is thus central to providing the research information management needs of an organisation and, indeed, to exposing (subject to permissions) that information driven from the CERIF-CRIS through organizational web pages. A CERIF-CRIS performs all these functions effectively and efficiently because of the formal syntax (information structure) and semantics (meaning) of CERIF.

However, the user may wish to access information on research at other organisations. CERIF provides the necessary interoperation capability. Thus the research information of other organisations (including that in the repositories and the associated organisational systems such as finance, human resources, project management) also can be accessed and used (subject to security, privacy and other rights and restrictions) together with that of the user's organisation.

CERIF-CRIS rely on the e-infrastructure. For reasons of flexibility, effectiveness and efficiency such a CRIS is best implemented using a SOA (Service Oriented Architecture). In this way the CERIF-CRIS forms the universal backplane to research information – and more generally provides the core services of the i-infrastructure.

1 Introduction

We describe in this section the concepts of e- and i-infrastructure, of CERIF-CRIS and discuss the EU view and conclude that (CERIF) is the key to providing a Research Organisation view of the e-infrastructure. Section 2 proposes the hypothesis, that a CERIF-CRIS can provide such a view for a research organisation. Section 3 expands upon how this is achieved, notably placing a CERIF-CRIS (and associated systems) in the context of SOA (Service-Oriented Architecture) particularly in a GRIDs environment. The clear outcome is the importance of metadata, and how a CERIF-CRIS can provide – as well as research in-

formation data – metadata to provide homogeneous access to other research information data sources. We conclude (Section 4) that a CERIF-CRIS implemented in SOA provides the ideal environment for a research organisation.

1.1 e-infrastructure, i-infrastructure and k-infrastructure

The e-infrastructure means different things to different groups of people. Some regard it as the communications network, some include data stores, processors, detectors and instrumentation. Some include the middleware which provides a homogeneous view to software systems over the heterogeneous physical infrastructure. Yet others include information systems necessary for the applications – including decision support systems – to operate. For the purposes of this paper the e-infrastructure refers to the physical infrastructure with its software to provide homogeneous access over the physical infrastructure (lower middleware) and the i-infrastructure refers to the information systems and the middleware to provide homogeneous access to information (upper middleware). It is possible to imagine a k-infrastructure providing knowledge-facilities over information systems including data mining to create hypotheses and expert advisor systems for decision support (Figure 1). These three layers correspond to the original UK proposal for GRIDs in an e-Science environment (Je99a).

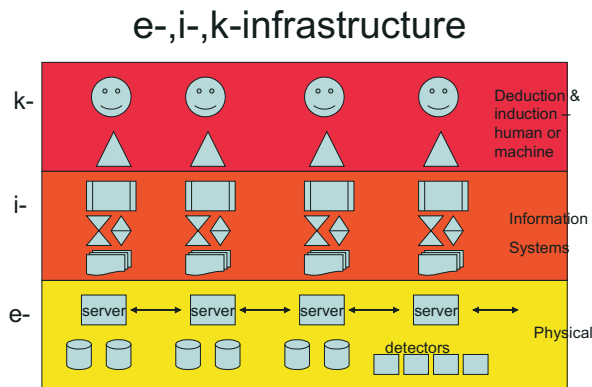


Figure 1: e, i and k-infrastructure

1.2 CERIF-CRIS

The CERIF-CRIS provides several kinds of functionality: it provides the information about research at an organisation for use by researchers (to find co-workers), research administrators / managers (to record and track research project progress), funding agencies (to assess results from funded research or to find reviewers), research organizations (to compare performance and make strategic decisions), entrepreneurs (to find ideas for exploitation) and the media (to explain the research to the public); increasingly it provides metadata for and access to scholarly publications stored in a repository (AsJe04) and research datasets and software stored in a different repository (Je04) to allow deeper investigation into the research.

Furthermore, it can provide access to information from other systems such as finance, human resources, project management, customer relationship management, stock control etc at the organisation (JeAs06a). In this way the CRIS provides primarily a view of the research domain but secondarily a permitted view into related information systems of the organisation to provide more detailed information such as financial transactions associated with a project or person or the CV and employment record of a person; it can act as the central authentication and authorisation resource, controlling user access to information and also controlling the workflow of e-processes; linked with this function it can provide directory services for query, email direction, internet phone, videoconferencing, messaging, calendar-matching and other purposes such as organisational or functional charts; it can be used to provide directly - or provide access to (via other systems) - the information to be displayed on web pages representing the organisation driven by or through the CERIF-CRIS.

The CRIS can interoperate using CERIF, and thus provide the portal so that a user of CRIS at organisation A – with associated repositories and information systems for finance etc at A – can access a CRIS at organisation B and thus its associated repositories and information systems as well.

To illustrate, let us start from the user point of view. She accesses her local CRIS. She obtains information on research appropriate to her query stored in the CRIS. She may also – via the CRIS – have access to the more detailed information in the associated systems such as the repositories and finance, human resources etc. Finally, she may – via the CRIS – access other CRIS (using CERIF for interoperability) which in turn may access their associated information systems (repositories, finance, human resources etc).

Such a system (with CRIS and repository components) has been in operation for some years at STFC in UK and others (e.g. UiB Norway) are completing building such an environment.

1.3 The EU View

Such a homogeneous research information environment has long been a goal of the EU (European Union); the latest statement of the vision is the ERA (European research Area). However, the initiative started with individual nation-states. In 1984 a pilot interoperation project was initiated between France, Italy and UK (IDEAS) (JeLaMiZaNuVa89) and extended in 1987 to the G7 countries (EXIRPTS) (NaJeBoLaVa92). The European focus came with recommendations from the conference of European Rectors' conferences (now European University Association) and the recommendations of the CREST working group of the European Commission Directorate for Research that the EU should have an interoperation standard. A group of national representative experts was convened; CERIF91 was the result succeeded by CERIF2000 and subsequent euroCRIS-managed CERIF updates. Full information on CERIF is available at www.eurocris.org. More recently euroHORCs has considered the requirement and a working group led by ESF (European Science Foundation) has reported: a key recommendation is that euroHORCs members should join euroCRIS and converge their systems to CERIF compatibility.

In parallel there has developed, from the world of digital libraries, a repository-based view of research information, starting with scholarly publications but – in some cases – extended to research datasets. Using (DC) (Dublin Core) as the metadata and (OAI-PMH) as the interoperation protocol, various systems have been constructed and tested e.g. (ePrints)

and (DSpace). A model framework for digital libraries in a GRIDs (e-infrastructure) environment has been proposed by the (DILIGENT) project and is being extended to research datasets in the (D4SCIENCE) project. More recently, the (DRIVER) project aims to integrate European repositories of scholarly publications. However, the problems with DC as metadata have been discussed (Je99), in particular the lack of a formal syntax and the lack of declared semantics. More recently (2007) this difficulty has been recognized within the digital library community and attempts have been made to overcome it by encoding DC in (RDF) (DCinRDF) which is an evolutionary approach towards the model proposed in (Je99b).

It remains the view of the authors that the use of CERIF as metadata to describe contents of repositories and the use of CERIF as the interoperation data model is more effective and efficient than the proposals from the digital library world because of the formality of the model and its accurate representation of real-world relationships.

1.4 CERIF is the Key

The key is CERIF: its standard and formal syntax (information structure) provides efficiency and effectiveness both as a storage format for retrieval and as an exchange format for interoperation. This allows reliable transfer of information for processing. Its formal semantics – normalised in CERIF2006 – ensure interoperation at the knowledge level such that the meaning of the information is both human- and computer-understandable. This extends the capability of CERIF-CRIS and provides reliability, dependability, scalability and maintainability because more of the information management is undertaken by the software.

2 Hypothesis

The GRIDs community distinguishes lower and upper middleware at the e- to i-infrastructure boundary. Here we add k-upper and k-lower middleware at the i- to k-infrastructure boundary. The hypothesis of this paper is that – in the domain of research – the lower and upper middleware of the i-infrastructure (i.e. upper middleware and k-lower middleware) and associated information systems should be provided by or through a CERIF-CRIS (Figure 2).

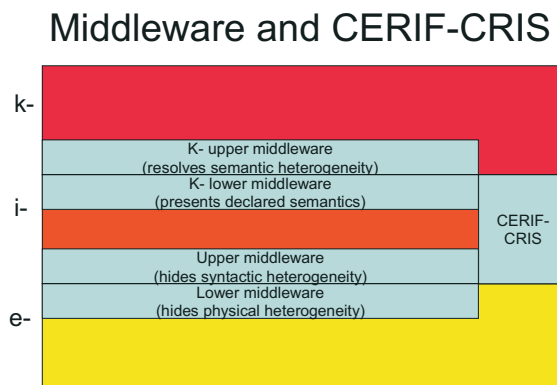


Figure 2: CERIF-CRIS in the infrastructure

This unifies the two concepts: that CERIF-CRIS provide the ideal central system to manage a research organisation and that the management of such an organisation – using a CERIF-CRIS - should be interfaced with the e-infrastructure. The corollaries are: that this architecture is optimal for the purposes of a research organisation; that this architecture interfaces optimally with the e-infrastructure, particularly in a GRIDs environment with ambient access using SOA (Service Oriented Architecture); that this architecture provides homogeneous access over heterogeneous CRIS, repositories of scholarly publications, repositories of research datasets and software and associated management information systems (e.g. finance, HR, project management) in a research organisation; that this architecture can provide a central data repository to drive many systems of an organisation such as directories, email addressing, web pages of organisation structure and functions, authority files for other systems such as finance especially when workflowed; that this architecture is open and expandable with evolving requirements of a research organisation; that this architecture allows interoperation (homogeneous access across heterogeneous research organizations) involving the CERIF-CRIS as the entrypoint not only to research information but also to repositories and associated management systems e.g. finance.

Elsewhere we have discussed and demonstrated the concept of the CERIF-CRIS as the central system for a research organisation (JeAs06a). Below we concentrate on how a CERIF-CRIS is interfaced to the e-infrastructure.

3 CRIS and SOA

We propose that the way forward for CERIF-CRIS is to be constructed using SOA principles. In this way it can integrate effectively with other components in the i-infrastructure and interface effectively to the e-infrastructure. Recently there has been a convergence of thinking from the GRIDs community and the SOA community with roadmaps being produced from respectively the Challengers Project (Challengers) and the 3S Project (3S).

3.1 SOA

A service provides a function that is well-defined (e.g. count of publications for a person) and a level of service defined by non-functional requirements (eg performance, reliability, precision, accuracy, rights, security, privacy and cost (if any)). Services can be (pre-) composed, or orchestrated statically or choreographed dynamically to form complete functional subsystems eg for update, retrieval, data exchange, data analysis (statistics), modelling (eg for what-if questions) and decision support.

The major advantage of SOA is that a program can assemble the services into a workflowed sequence to achieve the required task, which in turn ensures higher data quality in the CRIS because of the reduced threshold barrier related to data input as seen by the user (JeAs06b). Moreover, the software can assemble replicate service copies for parallelism to increase performance (e.g. searching several repositories in parallel) or resilience (e.g. having a service ready to operate if another fails). In advanced systems this composition could be done automatically during execution based on information exposed by each service during execution on its non-functional parameters (represented as metadata). This achieves so-called genetic programming (i.e. self-modifying software) but at the modular level rather than the programming statement level (Figure 3).

3.1.1 SOA and CERIF-CRIS

In the CERIF-CRIS world there is a requirement for services covering bulk load and update of data, and for management information. The latter could be covered by services representing the relational operators and associated functions such as count, sum, average. With benefit services could also include basic statistical functions and graphical display. Further services could include reporting tools, advanced statistics and data mining.

In addition, for interoperation, services for data exchange – including both data converters and communication - are required and can be combined with the other operators to provide homogeneous access over heterogeneous CRIS and associated systems.

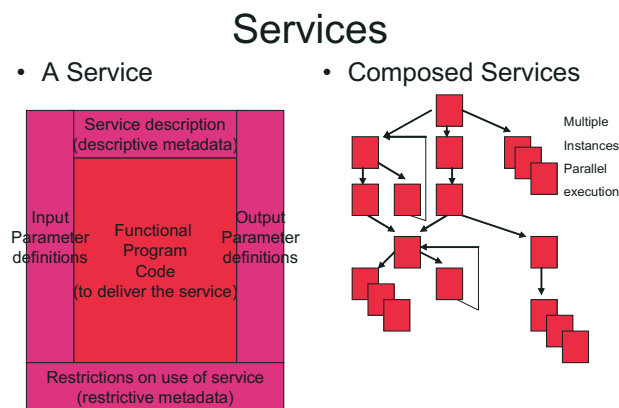


Figure 3: Service-Oriented Architecture

3.2.1 Example

The benefits of SOA can perhaps best be explained by example. Let us assume the end-user requires as output a set of graphs showing the number of output publications by department by year for the last 5 years across a university in Norway and a university in Germany. The services required are as follows:

1. discover the CRIS (perhaps with associated repository of full-text scholarly papers) for each of the two universities
2. from a web-form compose a query of the form 'select * from result_publication where type = 'peer-reviewed' and date > 2002-00-00; order by date; count by year
3. dispatch the query to the targets as discovered
4. translate this query into the target query language of the remote CRIS in each case remotely (i.e. either this service is provided locally or mobile code is dispatched there) where the target is either a CERIF-CRIS or a CERIF-wrapped CRIS;
5. execute the query and generate the (sorted) results including the count
6. dispatch the results from each of the target CRIS back to the end-user

7. at the end-user location the two result sets are used to generate the graph using local controls for the scaling, colours etc

Each of these steps would be a service and each is composed of more atomic services. However, it is possible to generate two or more instances of service 1 to get parallel and faster discovery. It is possible to generate two instances of services 3, 4, 5, 6 so that the two queries on the two target CRIS are executed in parallel. If the query on one fails because of CRIS unavailability, the system could generate a query (wrapped using OAISTER by OAI-PMH) and send it to the repository associated with the failed CRIS and retrieve appropriate publications from there – an example of alternative strategy to achieve the desired result. Finally, the local system of the end-user could provide alternative services for graphical representation such as EXCEL or a graphics package – thus demonstrating the flexibility and choice available with a SOA.

3.2 Metadata

In order to effect this composition each service (and within certain services the associated datasets or software components) must be described by metadata which covers the functional characteristics and the non-functional characteristics (i.e. descriptive and restrictive metadata), by navigational metadata for location and by the schema metadata to control integrity. This structured classification of metadata was introduced in (Je98) and elaborated in (Je00). Such a metadata description may be regarded as a (more elaborate) object-oriented class, knowledge engineering frame or functional programming signature (Je04).

Schema metadata constrains the data, or service in a formal way to ensure integrity (correctness).

Navigational metadata provides the route to access the data or invoke the service; commonly it is an (extended) URI.

Associative metadata comes in three kinds: descriptive (like a library catalogue card) which describes the data or service in a computer-readable and computer-understandable form; restrictive (like a conditions of use statement) which controls access, manages privacy and security, states any charges for the service or data access and generally ensures continuity of the business by availability; supportive which is metadata pertaining to a whole domain (not just a dataset or service) and which supports the services to knowledge level utilizing dictionaries, thesauri or domain ontologies.

3.2.1 Metadata and CERIF-CRIS

The metadata scheme outlined above can apply to the CRIS environment. CERIF provides both data structured as information (as used in, eg, decision support) and also as metadata to provide the contextual linking to e.g. repositories of scholarly publications or a finance system.

In the case of CRIS databases, the schema metadata describes the data structure in a formal way, the navigational metadata provides the URI for access and the associative metadata provides additional information. Descriptive associative metadata can be used to describe a particular logical record in the CRIS with primary attributes used for searching, indexing and statistical operation control (e.g. count, average, sum). Restrictive associative metadata can be used to restrict access to a particular logical record in a CRIS, either with

security or privacy or by making the information available subject to a charge or a disclosure of accessor and/or planned usage. Associative supportive metadata is used to provide the semantic interpretation of the database, usually in the form of a domain ontology with the attribute values, their inter-relationships (e.g. super and sub-terms) and their meaning – optionally in multiple languages.

For CRIS services, the schema will constrain the integrity of the service and thus ensure appropriate integrity, correctness and completeness in operation. The navigational metadata provides the URI to access the service and execute it. The associative descriptive metadata describes the service such that it can be discovered and utilised including in composition. The associative restrictive metadata describes the ‘conditions of use’ of the service and may include a price, but also could describe contractual rights, service level conditions or security / privacy constraints. The associative supportive metadata provides semantic clarity to assist in the utilization of the service and may be multilingual.

3.3 SOA-CRIS: The Way Forward

SOA is commonly associated with the evolution from WWW (World Wide Web) and web-services to GRIDs and GRID-services. Thus the adoption of SOA by CRIS will permit their presence on and participation in GRIDs and consequently their increased usage by all classes of users. The presence of CRIS in a GRIDs environment has been discussed in previous CRIS Conference papers by the authors (Je04) and in euroCRIS seminars.

The extension to hyperactive objects which are self-managing and dynamically work-flowed has been discussed in the context of grey literature (JeAs06c). This concept has implications beyond conventional service-orientation and relates to the SOKU (Service-Oriented Knowledge Utility) concept (NGG06). Thus it is possible to move incrementally from a SOA-based GRIDs environment for CRIS to a more autonomic, SOKU-based environment with excellent properties in scalability, resilience, performance, dynamic response to requirements and extensibility.

4 Conclusion

It is clear that SOA designed CERIF-CRIS can provide the data, information and knowledge backplane of the e-, i- and k-infrastructure for the research and development domain. The work to be done is as follows:

1. Metadata
 - a) Provide a strawman proposal for metadata for services, objects based on schema / navigational / associative (descriptive, restrictive, supportive)
 - b) Gain general agreement
 - c) Test in use-cases
 - d) Revise as necessary
 - e) Standardise (ETSI, W3C)
2. Services
 - a) Provide a strawman proposal for CRIS services based on relational algebra and additional functions
 - b) Gain general agreement

- c) Test in use-cases
- d) Revise as necessary
- e) Standardise (ETSI, W3C)

euroCRIS should be active in these tasks to ensure CERIF-CRIS well-represented in the ICT infrastructure.

References

- (3S) <http://www.eu-ecss.eu/>
- (AsJe04) Asserson, A.; Jeffery, K.G.; 'Research Output Publications and CRIS' in A. Nase, G. van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 29-40
- (CERIF) www.eurocris.org/cerif
- (Challengers) <http://challengers-org.eu/>
- (D4SCIENCE) www.d4science.org
- (DC) <http://dublincore.org/>
- (DCinRDF) <http://www.ukoln.ac.uk/metadata/dcmi/rdf-values/>
- (DILIGENT) <http://www.diligentproject.org/>
- (DRIVER) <http://www.driver-support.eu/en/about.html>
- (DSpace) www.dspace.org
- (ePrints) www.eprints.org
- (ePubs) http://epubs.cclr_Hlt126331628_Hlt126331629cBM_4_BM_5_.ac.uk
- (Je98) Jeffery, K.G.: 'Metadata' Invited Paper CRIS98 Conference, March 1998, Luxembourg.
- (Je99a) Jeffery, K.G.: 'Knowledge, Information and Data' September 1999 Paper submitted to Director General of Research Councils available at <http://www.semanticgrid.org/docs/KnowledgeInformationData/KnowledgeInformationData.html>
- (Je99b) Jeffery, K.G.: 'An Architecture for Grey Literature in a R&D Context' Proceedings GL'99 (Grey Literature) Conference Washington DC October 1999
- (Je00) Jeffery, K.G.: 'Metadata': in Brinkkemper, J.; Lindencrona, E.; Solvberg, A. (Eds): 'Information Systems Engineering' Springer Verlag, London 2000. ISBN 1-85233-317-0.
- (Je04) Jeffery, K.G.: 'The New Technologies: can CRISs Benefit' in A. Nase, G. van Grootel (Eds) Proceedings CRIS2004 Conference, Leuven University Press ISBN 90 5867 3839 May 2004 pp 77-88
- (JeAs06a) Keith G. Jeffery, Anne Asserson: 'CRIS Central Relating Information System' in Anne Gams Steine Asserson, Eduard J Simons (Eds) 'Enabling Interaction and Quality: Beyond the Hanseatic League'; Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen, May 2006 pp109-120 Leuven University Press ISBN 978 90 5867 536 1
- (JeAs06b) Keith G. Jeffery, Anne Asserson: 'Supporting the Research Process with a CRIS' in Anne Gams Steine Asserson, Eduard J. Simons (Eds) 'Enabling Interaction and Quality: Beyond the Hanseatic League'; Proceedings 8th International Conference on Current Research Information Systems CRIS2006 Conference, Bergen, May 2006 pp 121-130 Leuven University Press ISBN 978 90 5867 536 1
- (JeAs06c) Keith G. Jeffery, Anne Asserson: 'Hyperactive Grey Objects' Proceedings Grey Literature 8 Conference, New Orleans, December 2006; TextRelease; ISBN 90-77484-08-6. ISSN 1386-2316; No. 8-06-X

- (JeLaMiZaNuVa89) K.G. Jeffery, J.O. Lay, J.-F. Miquel, S. Zardan, F. Naldi, I. Vannini-Parenti.: 'IDEAS: A System for International Data Exchange and Access for Science'. Information Processing and Management Volume 25 No 6 pp703-711, 1989.
- (NaJeBoLaVa92) Naldi, F., Jeffery, K.G., Bordogna, G., Lay, J.O., Vannini-Parenti, I.: A Distributed Architecture to Provide Uniform Access to Pre-Existing Independent, Heterogeneous Information Systems, RAL Report 92-003
- (NGG06) Future for European Grids: GRIDs and Service Oriented Knowledge Utilities: Vision and Research Directions 2010 and Beyond; NGG Group January 2006
- (OAI-PMH) <http://www.openarchives.org/pmh/>
- (OAIS) <http://ssdoo.gsfc.nasa.gov/nost/isoas/>
- (RDF) <http://www.w3.org/RDF/>

Contact Information

Prof Keith G Jeffery

Science and Technology Facilities Council
Rutherford Appleton Laboratory
Harwell Science and Innovation Campus
Chilton, Didcot, Oxfordshire
OX11 0QX UK
Email: keith.g.jeffery@rl.ac.uk

Anne Asserson

University of Bergen
5020 Bergen, Norway
Email: anne.asserson@fa.uib.no