Knowledge Exchange, Danish Library Agency
H.C. Andersens Boulevard 2, DK-1553 Copenhagen V, Denmark
Tel. +45 33 73 33 73, Fax +45 33 73 33 72
office@knowledge-exchange.info, www.knowledge-exchange.info

Knowledge Exchange

February 2007

Institutional Repositories Workshop Strand Report
Strand title: Exchanging Research Information

Matthias Razum
Ed Simons
Wolfram Horstmann

1 Executive Summary

The report gives an overview of the present state of discussion as it was experienced during the workshop, provides orientation in the field and highlights possible next steps to be taken. It is primarily written for the organisations working together in the Knowledge Exchange network but also for experts in the field who can identify challenges and opportunities for collaboration.

The objective of the strand "Exchanging Research Information" was to bring together CRIS (Current Research Information Systems) and OAR (Open Access Repositories). Both applications deal with a specific segment of the *academic information domain* – notably the specifications, products or outcomes of academic research. Substantial commonalities exist between the two. Rooted in different units of the university (research administration vs. library) they, however, also have their individual characteristics: CRIS primarily have an institutional scope and are mainly referring to *context* of research whereas OAR are referring to *content* of research and are per definition internationally oriented.

Given their affinity, achieving interoperability between CRIS and OAR is desirable and will benefit all parties involved, including the researchers. A joint approach will avoid double input and management of redundant data as well as redundant services and processes and will both enhance the efficiency and quality (mutual enrichment) of the services offered by CRIS and OAR to their users.

Looking at the current situation in the KE-countries, significant differences can be noticed between Denmark (unified system), the Netherlands (strong national CRIS solution METIS with first integration with repositories), the UK and Germany (heterogeneous landscapes of institutional and subject-based repositories and less standardized CRIS). Successful integrated solutions can be found at an institutional or subject-based level, but integration becomes less probable moving towards complex landscapes at the national and supra-national level.

As a specific consequence of this situation, *ad hoc* technical developments to support interoperability between CRIS and OAR *at large scale* are currently recommended to be highly focussed on a specific entity in the academic information domain (e.g. managing the full-text-link of a research paper). As a broader consequence, a sustainable and optimal solution for the combination of CRIS and OAR *at large scale* requires a thorough analysis and specification capable of representing the heterogeneity of the two respective landscapes. It requires a flexible, service-oriented approach based on an integrated institutional policy concerning the academic

information domain, and targeting both organizational aspects (taking account of business processes and services) and technical aspects (implementation of service oriented architecture).

One step to take in this respect – and the first part of a follow-up activity of the strand – would be the delineation of the academic information domain, and notably the part within the academic information domain that is covered by CRIS and OAR. This would involve an analysis of the information elements (*entities and attributes*) and the workflows and services involved with CRIS and OAR. Another, parallel, step – and the second part of a follow-up activity of the strand – would focus on *ad hoc* technical development for managing a specific entity in the academic information domain in both CRIS and OAR. Once this work is done and the results of both steps are integrated, the definition of an optimized services model, integrating CRIS and OAR becomes more feasible and can be based on the principles of reuse of services (also on a supra-institutional level) and proper ownership of data. Such a new services model may have an impact on existing technical solutions (decomposition of systems) and even on organizational units (restructuring of business processes and workflows).

The strand recommends to the KE Board to initiate, and provide ways of funding, for the follow-up activities identified above. An adequate instrument is a sequel of the workshop. The overall goal should be preserved but the strand should be split-up in two groups: one group is working on policies and services in the academic information domain with experts for high-level and middle-level research management, and the other group is working 'hands-on' to build a demonstrator for achieving interoperability between CRIS and OAR for a specific entity in the academic information domain. Responsibility for a joint report should ensure mutual exchange.

2 Summary of recommendations

2.1 Work against the background of an integrated information policy and management in the institution, concerning the Academic Information Domain

- Integration and optimizing of business processes and workflows
- Institutional policy for integrated management
- Researcher-centred approach
- Open for re-use outside of the institution
- Apply a service oriented architecture

2.2 Think against the background of a distributed architecture

- 'Contracts' between data providers (e.g. service-level agreements)

2.3 Apply a service reference framework (e.g., e-Framework by JISC and DART)

- Start a follow-up activity within the KE framework to delineate the entities, attributes, services and workflows of the academic information domain as far as CRIS and OAR are concerned, as a concrete follow-up to the strand's discussion.

2.4 Work out an operational example of a technical solution

- Start a second and parallel follow-up activity to develop – as an example and demonstrator - a technical solution for the management of a specific entity within the academic information domain (e.g. the full-text-link of a research paper), common to both CRIS and OAR.

3 Discussion (including recommendations and items of interest)

3.1 Setting the Problem

3.1.1 The Academic Information Domain (AID)

Within the overall setting of information supply and management of academic institutions, a distinction can be made between information elements (entities and attributes) which are specific or typical for the academic work processes (research and education) and those which are more generic or primarily belong to other domains of work. An example of the latter are name and appointment information of authors (academic staff members), which belong to the personnel administration domain.

Information elements specific for the academic process, together with the workflows and services managing and using them, constitute what could be called the *Academic Information Domain* (as distinguished from e.g. the *Personnel Information Domain* or the *Financial Information Domain*). Examples for entities of the AID can be found in section 3.3.4.

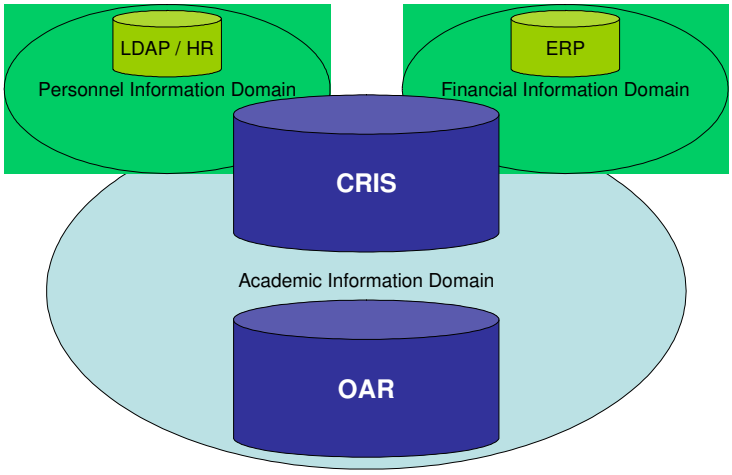

Figure 1: The Academic Information Domain (AID)

3.1.2 Introduction to Current Research Information Systems (CRIS)

CRIS are – mostly relational database – information systems containing an extensive set of metadata covering the various aspects of research information. They are in use at universities, research institutes, and/or governmental bodies and were initially developed for administrative purposes, notably:

- Reporting of research

- Assessment and evaluation of research
- Management of research

In recent years, CRIS have evolved towards web-based systems with a more *researcher-centric* focus, putting the added value for researchers, e.g. the international exposure of their research, as a priority[1].

Given the extent (and complexity) of the metadata involved, CRIS are based on well-worked out data models: mostly the CERIF-model (Current European Research Information Format), a European standard worked out by EuroCRIS [1] and recommended by the EU, or CERIF-compatible models.

As for the main characteristics of nowadays CRIS (and the metadata in them); they are:

- Extensive *(covering research activities, organizations, researchers, in- and output of research)*
- Detailed *(metadata broken down to their most detailed "atomic" level)*
- Formalized *(i.e. "schemed", e.g. enumerated lists, dictionaries, thesauri)*
- Logically structured and normalized *(a consequence of their RDB-nature)*

3.1.3 Introduction to Open Access Repositories (OAR)

A convenient definition for an Institutional Repository is a "digital collection capturing and preserving the intellectual output of a single or multi-university community" [2]. Many repositories are managed at a single-institution level. Other setups include consortial or multi-institutional basis. Beside Institutional Repositories, discipline-specific repositories are common, especially in physics and economics.

Open access offers free and unrestricted access through the internet to primary scholarly material (e.g., datasets) and secondary scholarly material (e.g., publications). OA archives or repositories do not perform peer review, but simply make their contents freely available to the world. They may contain unrefereed preprints, refereed postprints, or both [3]. There is growing international momentum in favour of institutional repositories. Increasing numbers of libraries are taking on the role of hosts for institutional repositories, becoming responsible for maintaining the intellectual heritage of their institution.

The Open Access and Institutional Repository community relies on a distributed architecture. Instead of maintaining a few central systems (which exist, e.g., ArXiV for physics and computer science), institutes and organisations are encouraged to host local repositories. This is due to differences in policies, workflows, scope of repositories, metadata requirements, the 'branding' of contents, and improved visibility of the organisation. Interoperability across distributed repositories is thus a major requirement for enhancing scholarly communication. This led to the formation of the Open Archives Initiative (OAI). OAI Repositories implement the OAI Protocol for Metadata Harvesting (OAI-PMH), thus enabling the dissemination and harvesting of metadata across repositories [4][5].

OARs currently broaden their scope and embrace more and more non-traditional publications (e.g. learning objects, multimedia assets, and datasets). In many cases, the actual content is not

[1] One could even say that "Current" in CRIS is therefore becoming less appropriate. The CRIS may very well include some completed (non-current) research activities, which leads to a more generic focus.

ingested into the repository, but resides in specialized systems that are designed for that kind of data. Examples are primary datasets or streaming media. OAR rather reference than copy these objects.

3.2 Observations

3.2.1 Differences

Scope

CRIS focus on institutional tasks. Aggregation or federation beyond institutional (or national) boundaries is rare. The maintained data has a strong relation to the institution (e.g. researchers, projects, budgets). Information pertaining external persons or organisations is usually not fully qualified. Discipline-specific CRIS on an international level do not exist. In some cases they exist on a national level (e.g. METIS in the Netherlands).

OAR focus on global awareness and discovery of scholarly information and communication. This requires the aggregation and federation above the institutional level. Researchers often prefer discipline-specific repositories to institutional repositories (e.g., physicists use ArXiV).

User groups

Two separate user groups have originally initiated and are still maintaining CRIS and OAR: research administrators and librarians. Typically, the two groups work in separate organisational units within an institution with little interdependency. Archiving interoperability is therefore not only a technical issue, but also an organizational task, which includes overcoming organisational barriers, exploring and specifying business processes.

Metadata

The *raison d'être* for OAR is providing the correct content (in most cases the full-text). Metadata is important for discovery and dissemination. Good metadata quality is desirable, but incomplete or not thoroughly checked metadata may be acceptable, if otherwise the content would not be accessible. OAR employ various metadata schemas driven by diverging object types, workflows, and applications. Dublin Core has been established as the smallest common denominator and is used in OAI-PMH, even though various other metadata schemata exist and are often employed in parallel by OAR. OAR store a snapshot of the descriptive metadata with the object at the time of its registration. Repositories will not update this snapshot, even if e.g., the author's name or affiliation changes later on.

CRIS metadata are referring to context. Typically, metadata is semantically rich and machine-understandable. A well-established metadata schema on the European level is CERIF. Metadata quality is pivotal for CRIS, as careers of researchers may depend on them. Exposure of research data including publications is a primary goal for CRIS, even though this is a recent development. Interestingly enough, CRIS typically care only about the researchers in their own organisations. Co-authors from other institutions are handled as simple entities with no affiliation information and practically no metadata. Therefore, this data is not as valuable for the OAR world as originally expected.

3.2.2 Commonalities

Both systems address the needs of academics. Even though the systems started from different backgrounds with differing user groups and scopes, they converge into a researcher-oriented

service, which includes objects and metadata resulting from the academic process. As such, they are not limited to research material, but cover the broader academic work context. Both systems maintain similar type of metadata with substantial overlap.

Future systems will integrate research material, educational resources, learning objects, e-Thesis, and datasets (objects within the AID) and relate them to administrative data like personnel, financial, and project information.

3.2.3 Interoperability

Usage scenarios for interoperability within the OAR landscape on the one side and between OAR and CRIS on the other side differ substantially. OAR typically harvest data from various other OAR, which are spread geographically and organisationally. In most cases, no authorization is required (anonymous harvests). Interoperability between CRIS and OAR focus on a limited number of well-known systems within the same organisation, which reduces the complexity of the task. On the technical level, a set of interfaces and exchange formats have to be designed and implemented. In fact, current CRIS have already many interfaces to other systems (financial, personnel, LDAP). The difficulties in integrating the two worlds mainly lie in the organisational and social barriers.

3.2.4 Non-exclusive Information Exchange

Information exchange needs to be non-exclusive. Often, OAR are broader in scope and contain additional object types that are not relevant to CRIS. On the other hand, CRIS sometimes include publications from researchers that were published before they joined their current organization[2]. CRIS may need to broaden their scope to include e-learning objects, multimedia assets, and datasets. This may be driven as well by a future shift in research evaluation and assessment that embraces non-traditional publications.

3.3 Exploration

It proved to be a prerequisite for the strand's discussion to come to a common understanding of each systems features and qualities. As already noted, the delineation of the AID is a prerequisite to interoperability. Hence, it proofed to be worthwhile to study the problem space of the AID by identifying its business processes before elaborating on formats and exchange protocols. Participants agreed on the importance of developing a joint vision of such scenarios, thus avoiding the risk of separate views on CRIS and OAR systems. Additionally, they stressed the need to identify rather generic processes that make sense internationally (or at least in the context of KE participants). Many issues have already been addressed and sometimes solved on a national level, especially in the Netherlands.

As a first step towards the delineation of the AID, the strand's participants explored one common business process (deposit mechanism / publication registration process). The idea was to concentrate on a simple usage scenario as focal point for exploration. The approach was to analyse the metadata requirements of both CRIS and OAR for this process, e.g. the full-text link of a research paper.

[2] This is not true for the Netherlands. In general, CRIS are exclusive for publications created during a researcher's appointment at a university.

3.3.1 Shared Metadata Requirements

The following list includes the metadata elements, which are relevant to both CRIS and OAR:

- (persistent) identifier
- publication metadata (bibliographic/descriptive MD, e.g., title, abstract, link to full-text)
- author and affiliation information
- keywords classification (based on controlled terms, i.e. formalized schemes)
- publication type (thesis, peer reviewed article, report, paper, etc.)

For detailed information on publication and author metadata, see workshop reports of the Research Paper Metadata strand and the Author Identification strand.

3.3.2 CERIF-specific Requirements

The following non-comprehensive list includes metadata, which is relevant mainly to CRIS:

- research unit
- research project
- research input (f.t.e. and €)
- scientific ranking
- roles
- funding body

3.3.3 OAR-specific Requirements

The following non-comprehensive list includes metadata, which is relevant mainly to OAR:

- full-text / content
- rights information
- collection information / sets / grouping
- provenance, e.g. publisher, repository
- version information
- manifestation publication type
- compound objects
- relational metadata
- technical metadata (mime-types, sizes, GDFR)
- preservation metadata

3.3.4 Schematic View

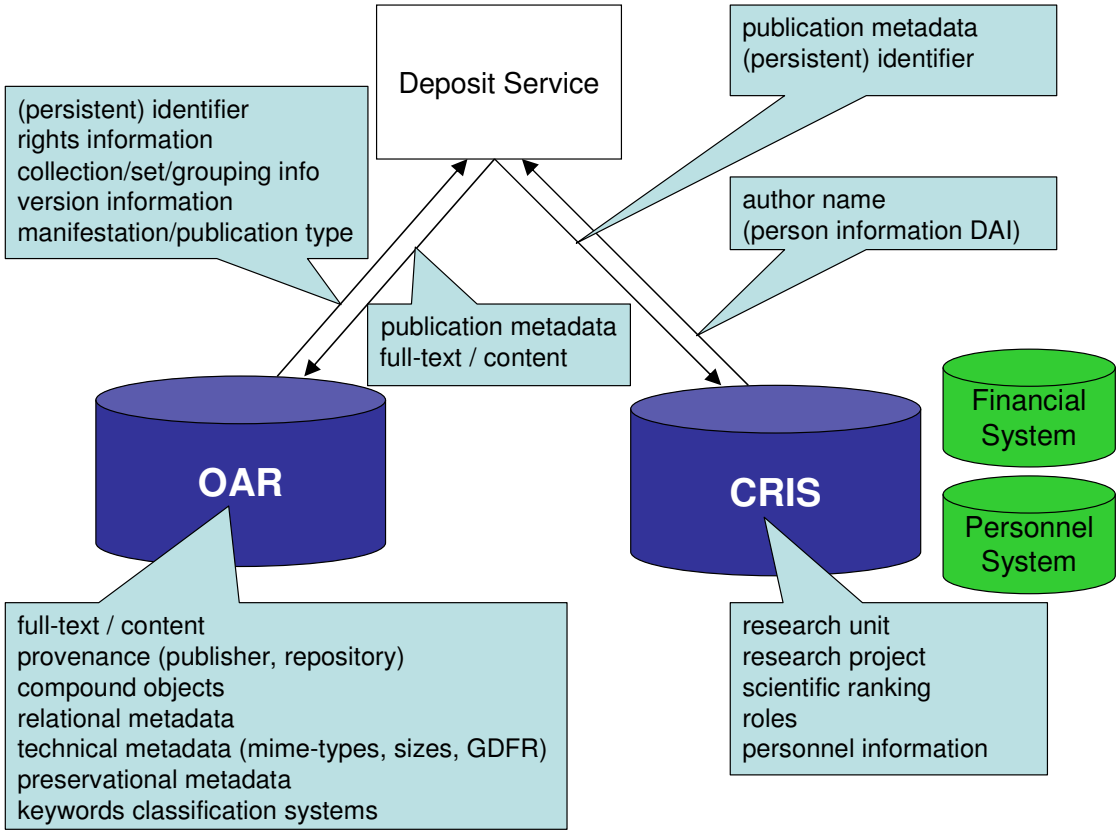

Figure 2: Data Flow between Deposit Process, OAR and CRIS. The Deposit Process can reuse data already available in other systems (see arrows pointing to Deposit Process) and provide publication metadata for CRIS and OAR.

4 Outcomes

Participants agreed that achieving interoperability between CRIS and OAR is desirable and would not only benefit research administrators and librarians as maintainers of these systems, but would create an added value to researchers as well, at least in avoiding double input of data. Out of the strand's discussions came that a clear and overall policy concerning the Academic Information Domain (AID) on the institutional level is a prerequisite for an optimal and sustainable solution of the exchange of information among or integration of CRIS and OAR. Such a policy should involve a clear delineation of the AID as well as an integrated vision and solution concerning the division of responsibilities and labour of the units and systems involved in the AID, among which are CRIS and OAR[3]. In a university, this policy should comprise a common concept (and workflow) for the CRIS and the OAR so that researchers do not feel to be approached by different departments for the provision of their data.

[3] Typically, within academic institutions an information policy often is worked out for the supportive information domains of an academic institution, such as personnel information or financial information, whereas an integrated policy for the AID, is still missing. All this while the AID concerns the core business of academic institutions and therefore could be expected to be in the centre of the information framework and policy (for a support of this vision from a research point of view see [6]).

To foster these policies and support uniformity on a European level, one of the follow-up activities of the strand's workshop could involve the AID-delineation mentioned above (definition of entities and attributes, workflows and services) as far as CRIS and OAR are concerned. The delineation process should describe in a first step the status of existing systems and solutions. In a second step, a working group might recommend a generally applicable model, thus fostering standardization and interoperability.

A second outcome of the discussions was that the registration and control of the information belonging to a given information domain is a matter of (the units of) that domain itself. Systems or services needing or using information from other domains should not register this information again but obtain (link, retrieve) it from the primary domain. Referencing information external to a system or service is preferable over copies, even though due to availability, reliability, and performance reasons, copies might be unavoidable. Keeping the (authority) information in the originating system improves the integrity, accuracy, and timeliness of data, thus advancing the overall data quality.

The outcome of a delineation of the Academic Information Domain would be an abstract and high-level description of a generic "container model", which would integrate all the information from not only CRIS and OAR, but other relevant data sources as well. Such a container model should not strive to be extensive and complex content or data model. Instead, it could be a virtual aggregation of resources stored in various existing systems (including OAR and CRIS). This approach is similar to the Danish model, which has proven to be quite successful.

In a following step, we recommend the formulation of an optimized services model, involving the identification of common services that are reusable by both CRIS and OAR systems. Eventually, these services may become relevant to other external systems as well. Services for author/person information and affiliations – in the meaning of authority files – would certainly be interesting (but probably hard to maintain). This optimized service model may require a decomposition of existing systems. A guideline for such a decomposition into services could be the "service reference framework", done by Liz Lyon and Andy Powell [7].

Participants agreed that due to the differing approaches taken in the KE countries, interoperability on a pan-national level is only achievable in an international cooperation. As multinational projects will further increase in number, local (e.g., universities) or national scope of both CRIS and OAR will no longer be sufficient. Bringing together the varying levels and areas of expertise from DEFF, DFG, JISC, and SURF would foster the development of a sufficiently generic, thus generally applicable solution.

As a follow-up, the following steps could be part of a funded effort:

- specification of the problem space / Academic Information Domain, including
  - identification of relevant business processes
  - development of a high-level information model
  - identification of relevant data sources
  - system decomposition
  - development of a reference architecture for distributed services
  - demonstrator focussing on a specific entity in the AID

The outcomes of this step should then be evaluated. If the results are promising, further activities might include the specification of

- data exchange formats and protocols
- implementation of adaptors to existing systems (e.g., personnel, finance)
- implementation of reference services (e.g. deposit mechanism)

5 Annexes

5.1 Supporting Information

Briefing Paper

- "Exchanging Research Information" by Wolfram Horstmann

Presentations

- "CRIS and OAI - An approach from the CRIS-side" by Ed Simons
- "The OAR Perspective on Interoperability" by Matthias Razum
- "CERIF 2006" by Geert van Grootel
- "Danish Perspective on OAR/CRIS" by Alfred Heller
- "Data Integration in Current Research Information Systems" by Max Stempfhuber

5.2 Referenced Materials

[1] EuroCRIS: European Association of CRIS-providers, http://www.eurocris.org

[2] Raym Crow: *The case for institutional repositories: a SPARC position paper.*

[3] Peter Suber, http://www.earlham.edu/~peters/fos/brief.htm

[4] The Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH), http://www.openarchives.org/pmh

[5] Carl Lagoze, Herbert Van de Sompel: *The Open Archives Initiative: Building a low-barrier interoperability framework*, Joint Conference on Digital Libraries 2001. Draft available from http://www.cs.cornell.edu/lagoze/papers/oai-jcdl.pdf.

[6] 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 of the 8th International Conference on Current Research Information Systems, Leuven, University Press, 2006, p. 109-120.)

[7] Andrew Powell, Liz Lyon: *The DNER Technical Architecture: scoping the information environment*, 2001.
http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/

5.3 List of Participants

Lead Coordinator:

Dr. Wolfram Horstmann
State and University Library Goettingen
whorstmann@sub.uni-goettingen.de

Moderators:

Dr. Ed Simons
Radboud University Njimegen
e.simons@bz.ru.nl

Matthias Razum
FIZ Karlsruhe
matthias.razum@fiz-karlsruhe.de

Knowledge Exchange Representative:

Johannes Fournier
German Research Foundation DFG
johannes.fournier@dfg.de

Participants DFG:

Dr. Jan Brase
German National Library of Science and Technology
jan.brase@tib.uni-hannover.de

Dr. Maximilian Stempfhuber
GESIS/Social Science Information Center
st@iz-soz.de

Participants JISC:

Simon Lambert
CCLRC
S.C.Lambert@rl.ac.uk

Richard Jones
Imperial College London
richard.d.jones@imperial.ac.uk

Participants SURF:

Drs. Ruud Bronmans
Royal Netherlands Academy of Arts & Science
ruud.bronmans@bureau.knaw.nl

Dr. Ir FWAM Miesen
Library of the Eindhoven University of Technology
f.w.a.m.miesen@tue.nl

Drs. Ing. Renze Brandsma
Library of the University of Amsterdam
r.brandsma@uva.nl

Participants DEFF:

Alfred Heller, Ph.D.
Technical Knowledge Center of Denmark
ajh@dtv.dk

Henrik Juul-Nyholm
University of Copenhagen
hjn@adm.ku.dk

Guest:

Geert van Grootel
EuroCRIS
geert.vangrootel@wim.vlaanderen.be