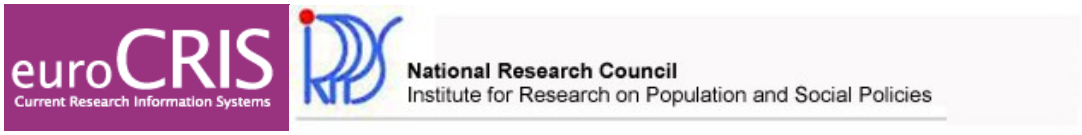

White Paper

State of the Art and Roadmap for Current Research Information Systems and Repositories

{Keith G Jeffery, Anne Asserson} euroCRIS, {Daniela Luzi} CNR IRPPS

Introduction

Research Information

There is much increased interest in, and demand for, research information. It is used by very many people including researchers, research managers, entrepreneurs and the media. It is necessary as part of the e-infrastructure of the European Research Area to support strategic decision-making, to assist cooperation and to stimulate wealth creation and improvement in the quality of life. It is the engine of the knowledge society.

CRIS

Since the 1960s CRIS (Current Research Information Systems) have been used particularly to manage research. In the 1980s CRIS from national funding agencies were interoperated in a prototype system. This led to the EC supporting the development of CERIF (Common European Research Information Format) as a data model for interoperation of CRIS, formally since 2000 an EU Recommendation to Member States. Since 2002 CERIF is maintained and developed under the auspices of euroCRIS (www.eurocris.org). Most CRIS provide a user-view (usually restricted for reasons of security, privacy and confidentiality) via the WWW which is toll-free open access.

Repositories

Repository systems also have a history. EPrints archives were developed in the 1990s. DSpace developed from EPrints and other repository systems followed. Many research institutions have an institutional repository of scholarly publications using these repository technologies. These usually are toll-free open access. Following on from the early SPIRES system (at CERN) the particular physics community developed arXiv in 1999 based on the Los Alamos National Laboratory system commenced in 1991. Some subject domains have a thematic repository (such as arXiv) and in the biomedical domain there are publisher-owned commercial repositories (author-pays rather than

access by toll). In recent years there have been EC-funded projects to link open access repositories of scholarly content such as DRIVER and DRIVER2 and more recently the OpenAIRE project aims to collect together research outputs from EC Framework 7 projects.

Bringing Together CRIS and Repositories

Various pressures bring CRIS and repositories together. In many countries evaluation of research is becoming important, to assure public funds are well-spent. Such requirements need to associate research output (including publications but also products, patents, dissemination at events) with persons, organizational units (university, faculty, department, cross-department centre, research group), projects (e.g. funded grants), facilities (e.g. large synchrotron radiation sources), equipment (particular experiment) for analysis. Repository systems usually do not collect this contextual information, being restricted to scholarly publication output with title, abstract, author and some other publication-related information, commonly based on DC (Dublin Core) metadata. On the other hand, some CRIS include within them the research outputs as objects (e.g. publications). Similarly, the context of a research output is important in evaluating its utility, quality and relevance to any repurposing. Context is usually recorded in a CRIS. CERIF provides an appropriate time-stamped structure so that provenance information is also available.

Rome Workshop May 10-11 2010

A workshop held in Rome under the auspices of euroCRIS and CNR 10-11 May 2010 brought together by invitation key experts in the technologies and their use. It provided a locus for the communities to understand and appreciate the requirements and architectures each was using. There were deep discussions on appropriate metadata, architectures, policies and best practice, all with a view to improving the utilisation and management of research information.

The Requirement

The Researcher

Researchers in general wish for recognition and for effective and efficient dissemination of their work and conduct of their research. Thus systems that produce automatically a researcher CV, bibliography or webpages are advantageous. Similarly systems providing instant access to relevant publications (white or grey), to relevant projects, to researchers in the appropriate domain, to funding sources and opportunities, to news about research, to relevant conferences and events are of great utility. Finally all such systems should integrate with the office and cooperative working environments and with institutional systems for administration and management such as finance, human resources, project management. Low administrative burden and once-only input of data should be warranted.

The Research Manager

Research managers exist both in research funding organizations and research-performing organizations. The latter include both primary research organizations and research supporting organizations such as those running large scale facilities, providing data from detectors or providing library services. Research managers in funding organizations wish to have systems to track research

programmes and calls in other organizations, to find reviewers for proposals, to keep track of proposals and success rates and to document research outputs to ensure value for money. Research managers in research performing and supporting organizations wish to have systems to manage the intellectual property of the organization, to track project proposals, success rates, project outputs, events and media communications and associated finances and human resources. They are likely to need cross linkages from their CRIS to institutional systems for management and administration. Of course they wish to be able to observe competitors and compare performance.

The Entrepreneur

One of the most difficult processes is the transfer of knowledge and technology from research to wealth creation and improvement in the quality of life. The techniques of knowledge and expertise transfer are complex. However, the dialogue between researchers and entrepreneurs may be assisted by well-documented research projects and research outputs. The usual problem is to recognize in the description of the research and its outputs the idea or product that could be used for innovation. However, the more research information that is available, the better the chance of achieving knowledge transfer. The achievement of knowledge transfer is recorded in patents, products and other registers.

The Media

Especially in the case of publicly-funded research, it is important to communicate information about research to the general public and this is usually done via the media. Much research activity and output is media friendly: witness the deep public interest in the current Icelandic volcanic activity, but also in environmental issues. Astronomical images are always popular as is information about research results and breakthroughs in biomedical science – especially new clinical procedures or new pharmaceutical products. The key is to feed the media channels with appropriate information.

The Policymaker

Especially in publicly-funded research, there are always difficult choices to make, especially concerning the prioritization of funding across domains of research. However, decision-making is assisted by verifiable information on the cost-effectiveness of research (usually measured by the outputs for a given cost and their quality), its impact on society (wealth creation and improvement in the quality of life), its effect on education and its effect on general knowledge in society.

The State of the Art

CRIS

CRIS have existed since the 1960s for the management of research activity. Both research funders and research performing organizations (including those supporting research through experimental facilities) utilize CRIS. CRIS were used to track research funding and activity, to catalogue research outputs, to document researcher CVs and bibliographies and to publicise research activity both for innovation and technology transfer but also as a public relations exercise. In the 1980s it was realized that in addition to managing activity in one organization access to CRIS of other organizations brought benefits (a) in looking for cooperative areas of research, (b) in comparing

metrics and (c) in locating reviewers for research proposals. Early prototypes demonstrated feasibility and led to the EC supporting the development of CERIF (Common European Research Information Format) as an EU Recommendation to member states. In 2002 the EC gave responsibility for CERIF to euroCRIS (www.eurocris.org) .

CRIS usually are based on structured database technology with considerable update activity as well as retrieval/reporting activity. Much of the retrieval is of the 'group by' kind: selecting all publications from a researcher or from a research department or from a project; selecting all researchers with >10 refereed publications per year; selecting departments with research income of >5 M€ per year. CERIF-CRIS, because the data model includes date/time stamping and roles, can provide the research context (project, persons, organizational units, funding, events, facilities, equipment) for a research output such as a publication leading to greater useful information for the end-user. It can also link together publications with products such as research datasets or software. A CERIF-CRIS can create a CV or recreate the state of an organization or project at any past time interval. Naturally CERF provides a mechanism for exchanging CRIS data or for providing homogeneous access over multiple distributed heterogeneous (or homogeneous) CRIS.

Repositories

Repositories (in the research domain) are used mainly to store scholarly publications although some repositories are built to store research datasets and software. Following from the CERN SPIRES database of particle physics publications, work at Los Alamos from 1991 and then Cornell led in 1999 to arXiv, a thematic toll-free open access repository for particle physics and related disciplines. EPrints archives were developed in the 1990s from which later DSpace was developed – all aimed at providing toll-free open access to scholarly publications in an institutional repository in parallel with subscriptions to publishers – with the aim of increasing access and utilization. During the 2000s publishers responded by providing open access to their individual commercial repositories on condition that the author pays for publication.

Repositories usually are based on querying by term-matching (as in classical information retrieval) over a semi-structured database. Usually there is no or little update after initial input. If repositories support OAI-PMH (Open Archive Initiative – Protocol for Metadata Harvesting) and some form of DC (Dublin Core) metadata (unqualified or qualified and utilising various subsets or supersets of the defined DC) then an OAISTER query can interoperate across multiple repositories.

Metadata

Traditionally metadata is defined as 'data about data'. In fact it is common for any data to be used and usable as both data and metadata. Consider a typical OPAC (Online Public Access catalog), the electronic equivalent of the library catalogue card system. An individual record is metadata for a user accessing a particular article but the collection of records is data for a librarian quantifying the number of articles with 'green widgets' in the title.

Metadata are used for many purposes:

1. Integrity: a traditional database schema ensures integrity of the instances under that schema;

2. Navigation: to reach the object of interest;

3. Associative-descriptive: to describe the object for discovery. This is the principal purpose of DC;

4. Associative-restrictive: to constrain use of the object of interest by rights including copyright, tolls;

5. Associative-contextual: to provide additional information associated with objects themselves associated with the object of interest to give context to improve understanding of the object of interest and its relevance to the query. Such metadata may include aspects of preservation/curation and provenance if it supports appropriate temporal representations;

6. Associative-supportive: this metadata relates to the domain of interest not the individual object and includes dictionaries, thesauri, domain ontologies and such resources that may be used to improve retrieval.

In general repositories include a small subset of the kinds of metadata described here, and the different kinds (with their different purposes and utilisation patterns) are mixed together. In contrast a CERIF-CRIS distinguishes clearly these kinds of metadata and supports the required temporal and role relationships to describe the complex real world of research .

Integration

There exist CRIS which contain within them the research output object such as the full text of a publication or video of a performance. At the other extreme recent work by the EPrints team has been focused on extending the EPrints datamodel to include relevant entities from CERIF to allow research evaluation. However, given:

a) The different access patterns to repositories and CRIS;

b) The different update patterns of repositories and CRIS;

c) The different (meta)data standards of repositories and CRIS;

there is merit in investigating how they can be used together for the benefit of research information provision. Indeed, there are many examples of CRIS interoperating with repositories. These examples range from the CRIS containing all the metadata and having just a unique identifier as foreign key to the primary key of the entry in the repository through separate input and maintenance of metadata in each but with shared unique identifiers to a repository holding a unique identifier to a research grant record in a CRIS.

However, the relationships between a research output publication and the various other entities in the research information domain are complex. There is increasingly a need to relate the research output publication to project(s), person(s), organizational units, facilities, equipment, events etc. These relationships have a temporal duration and include a role (such as personP *is author of* publicationX). Instances of entities involved in such relationships can take different roles at different time periods, and a unique instance of an entity may take different roles at different (or the same) time periods with respect to a unique instance in another entity. Example: personP *is manager of*

equipmentE; personP *is user of* equipmentE – this could be synchronous or asynchronous.

This indicates that for most purposes using research information simple primary key/foreign key (unique identifier) relationships are insufficient but the multiple temporal and role relationships between instances of entities are necessary. This leads inexorably to the need to store the (meta)data in the CRIS (which can sustain the richer data structure) and link to the repository to access the research output object. If OAI-PMH/OAISTER access is required for the repository then the appropriate subset of CERIF can generate DC and store it in the repository so ensuring consistency between the two system components within the totality of a research information system.

The Roadmap

The Target

It is assumed that the target is for the end-user (whether a researcher, research manager, entrepreneur/innovator or the media) to be able to find, select, integrate and display research information (where appropriate with other related information) at any place, anytime and in an appropriate form.

The Foundations

The roadmap, following the European e-Infrastructure Forum Report of 2010-04-28, assumes evolution towards an e-infrastructure of high-speed networking with data stores (guaranteeing curation), computation servers and detectors, managed by middleware in a GRIDs or CLOUDs configuration so that the underlying hardware resources are virtualized for the application software and the end-user.

It assumes also that applications will be encoded as services interfacing with the e-infrastructure through standardized interfaces for discovering resources, negotiating for use, using them (with monitoring for performance, security, privacy and if necessary payment) and appropriate logging and accounting.

It assumes the preferred end-user access channel will be WWW with a W3C-compliant browser.

The Steps

The major steps of the roadmap are as follows:

1. 2010: agree standard metadata format for research information usable by services, for interoperation (including wrapping legacy systems) and for linking with external systems as necessary;

2. 2010: liaise with European e-Infrastructure Forum (Ee-IF) and e-infrastructure Reflection Group (e-IRG) to agree the relationship between research information (including scholarly publications) and the output from research projects, facilities and equipment (including simulation on supercomputers);

3. Liaise via ESF with euroHORCs and via EC with CORDIS to ensure the standard metadata

meets requirements of research funding agencies;

4. 2010: initiate work on standardising metrics for the evaluation of research based on the agreed metadata standard in (1);
5. 2011: finalise standard metadata format;
6. 2011: agree standard services required for managing research information: these might include generate CV for person, generate bibliography for person or organizational unit, generate report of external research funding by project for an organizational unit, generate report of all current projects concerning <subject term> across Europe etc ;
7. 2011: liaise with Ee-IF, e-IRG, via ESF with euroHORCS and via the EC with CORDIS on the proposed services;
8. 2012: finalise standard services
9. 2012: prototype of standard services and standard metadata operating over e-infrastructure
10. 2012: evaluation of prototype
11. 2013: improvement of prototype to production status;
12. 2013: set up methodology for maintenance, development, evolution of standards for metadata and services.

The Developments Required

Metadata and Standardisation

Additional R&D work is required around the process of reaching agreement on the metadata standard. This involves particularly development of the semantic layer to ensure machine-machine interoperability without human intervention.

Services and Standardisation

Much additional R&D work is required to standardize services and to assure their interfaces through appropriate metadata sets. The information metadata can be handled by (developed) CERIF but the metadata to describe the services to allow automatic discovery, composition and execution (with dynamic reallocation including replication and parallelism) to meet a user requirement needs extensive R&D. The requirement places demands on the system beyond the usual web-services environment (UDDI, WSDL, BPEL4WS).

Migration to the new Architecture

Existing systems for research information in general do not interoperate, are not built on a common e-infrastructure and are insufficiently automated in system management and administration. It will be necessary to classify existing systems and for each class map a migration strategy to the new target environment. This is a complex and detailed task since it includes both information and processing environments, each with its own syntax, semantics, assumptions and business rules.