Národný informačný systém podpory výskumu a vývoja v SR
– prístup k elektronickým informačným zdrojom

NISPEZ

Informačné systémy o vede

Integrácia pre otvorený prístup k vedeckým výstupom

Zborník z medzinárodnej konferencie

CVTI SR, Bratislava 2. apríla 2014

CVTI SR, Bratislava 2014

Podporujeme výskumné aktivity na Slovensku / Projekt je financovaný zo zdrojov EÚ

Repositories, CRIS and CRIStin: the story so far

Anne Asserson

University of Bergen, Nórsko

anne.asserson@fa.uib.no

pracuje v univerzitnej knižnici Univerzity v Bergene, Nórsko. Má na starosti správu digitálnych systémov a služieb s dôrazom na výsledky výskumu.

Abstract

The management and utilisation of Research information has been evolving over many years. The paper traces the major threads of activity in repositories and CRISs, characterise them and discuss their usage. Further the paper considers the evolution of the technology to match the changing requirements of users for research information and indicate future directions.

Repositories

Repositories are ICT systems to store objects. Usually repositories refer to stored scholarly publication objects, although increasingly other objects – such as datasets – are stored in repositories.

Purpose

There are two basic stored objects in repositories: metadata and data. Typically the latter is the full text (or increasingly hyperlinked multimedia) of the scholarly article. The former has been subject to different influences namely (a) the librarian, cataloguing, community where the standard used internationally is MARC and (b) the web page community where the standard used internationally is DC (Dublin Core). The advantage of MARC is its relative richness compared with DC, its compatibility with library catalogues of non-electronic objects and its ability to hold all the attributes required for a full bibliographic reference. The advantage of DC is its simplicity and ease of use.

Experience of Utilisation

Systems for cataloguing scholarly articles have been in existence for many years; commonly researchers (or the library on their behalf) had handwritten and then punched cards with the bibliographic information of their publications later becoming electronic records in files. However, it was not until the advent of WWW that repositories with metadata and the full article appeared in any significant numbers.

Typically an institutional repository today will hold ~15% of the scholarly output of the institution as full text of the articles and a higher percentage as metadata recording the existence of the article. Some repositories hold only published (i.e. peer-reviewed) scholarly articles whereas others include all research output – suitably distinguished. The repositories are 'green OA' that is the articles are either the version as published by the publisher (some publishers insist on this, others prohibit it) or the final article as submitted for publication after peer review and editorial corrections. The business model remains the traditional one of library subscriptions for publication channels (journals, conference proceedings etc.)

Overwhelmingly the repositories use DC as the metadata standard although many repository systems have extended DC in divergent and incompatible ways thus making interoperation difficult. Retrieval is usually by text string search over all the metadata although systems where the search is restricted to particular elements (attributes) exist. The result is (compared with classical library catalogue systems) rather poor relevance and recall (precision).

The low degree of 'fill' of repositories can be ascribed to several factors: (a) the threshold barrier to input the metadata and load the article; (b) pressure and restrictions from commercial publishers through copyright and other legalistic mechanisms; (c) lack of perceived benefit. It has been demonstrated that articles in an open access institutional repository attract more citations than those not in the repository (Harnad, Swan 2009). This provides a refutation of (c). The factor (b) is significant; publishers have threatened authors legally if they deposit in an institutional repository and researchers clearly wish to have their work published in high-impact publication channels where their peers can see their work. Mandates by research funders and research institutions have provided researchers with protection and increased significantly the deposition of articles in 'green' repositories (Harnad, Swan 2009). although continued legalistic threats and legalistic uncertainties by publishers hinder progress. Recent systems to assist a researcher in managing their bibliography (Mendeley, ResearchGate, LinkedIn) have complicated the issue since they tend to store only the bibliographic metadata and provide a pointer (URL) to the scholarly article. Recent developments have provided a way forward – namely 'gold OA' where the publishers extract payment from the author / author institution for publication and then make the article available open access on the publisher website. Depending on the licence, it may be possible for a researcher (or an institutional library) to download a copy of the article for their own institutional repository. Initial estimates indicate that for a high-output research institution 'gold' OA costs three times as much as 'green'. However, the publishers are pushing this model singly and recently the UK government accepted this as the preferred model (Finch

2012) The (a) factor is also significant; the solution is to collect the metadata as available within the workflow of all the users (researcher, librarian, publisher) progressively and to input only once without any re-keying.

OAI-PMH is used as the interoperation protocol allowing harvesting of metadata from any compliant repository. The syntax of DC elements from different repositories can be combined together as long as the simple element structure is followed; the use of repeating groups or extended namespaces frustrates interoperation. Multilinguality and differing semantics (meaning of the lexical string) makes interoperation without much human intervention impossible.

The result is that using a repository –and through it a window to other repositories - requires much user interaction to filter the results manually to the required subset to meet the retrieval request intended rather than that expressed by the limited syntax and semantics. The lack of fill of repositories means that the end-user retrieval result is incomplete.

Management Information

A library online catalog system, especially when cross-linked to others – provides a basis for management of research information concerning scholarly output in the form of publications. Publishers also can utilise their catalogs for a similar purpose. The publication indexing services from Thomson-Reuters and Elsevier provide a reasonably comprehensive cover of scholarly publishing (there is a bias towards STM (Science, Technology, Medicine) and English-language) including citations. However, in these systems the recording of author names is variable (partly depending on the editorial style of the publication channel) and the recording of author institution inconsistent. This means that assignment of publications to persons and organisations is less accurate than is desirable. Attempts by Thomson-Reuters and others to assign author IDs and organisational IDs have been partly successful (although some ambiguities still exist particularly in evolving author names or evolving institutional structures and names). The recent ORCID initiative started with IDs for researchers (one step away from the role of authors) but realised quickly that much more metadata was required to allow disambiguation and so now is recording much of the information required by the CERIF model.

Increasingly there is a requirement – especially from research funders but also from research institutions managing their research portfolio) to relate scholarly articles to research funding (grant or grants) as well as to person(s) and organisation(s). Furthermore, there is a requirement also – when appropriate - to relate the publication to particular research facilities and/or equipment. These metrics are used for benchmarking institutional performance and for research funders to assess the research performance against funding for researchers and institutional organisations. It is clear that repositories using DC for metadata are incapable of meeting these new requirements for management information on research. (Jeffery, Asserson 2006)

In the last few years there has been a demand from governments to justify the research expenditure on the basis of impact of the research on society (wealth creation, improvement in

the quality of life) rather than the outputs. This new requirement is way beyond the capability of repositories.

CRIS

Purpose

The purpose of a CRIS is to record research information for the purposes of management and utilisation by a funder, research institution or researcher and for utilisation by innovators, entrepreneurs, the media and the public.

The development of the CERIF model by an expert group of national representatives convened by the EC was a major step forward for harmonisation and interoperability. CERIF91 was a 'flat' metadata scheme not unlike the (much later) DC; the difficulties were soon realised and the expert group reconvened to define CERIF2000 which – with subsequent evolution – provides a formal syntax and declared semantics for CRIS today. CERIF is used in 43 countries and is the national standard for research information in 10.

Experience of Utilisation

The period before 1980 is characterised by individual batch processing systems of the funding organisations. These systems recorded awarded (sometimes proposed) research project grants usually with identifier, title, PI (principal investigator) and institution of the PI together with the amount awarded. Sub-records recorded the financial transactions.

Subsequently the funding systems became more sophisticated including abstract of the research grant and keywords. This led to the requirement for a standard for interoperation and CERIF91 was developed. Subsequently the requirement was extended to cover research outputs and for flexible relationships between multiple funders, research organisations, persons and research grants / projects. This led to the development of CERIF2000. The relationship between research grant and associated postgraduate awards was recorded and the relationship to research facilities and equipment; among other requirements these have driven the evolution of CERIF to the current version.

CRIS are in widespread use and providing day-to-day management information for research funders and research institutions. Increasingly CRIS are becoming CERIF-compatible (a) to permit interoperation but more importantly (b) because the model reflects accurately the real world of research information. Several commercial companies offer CERIF-compatible CRIS and both Thomson-Reuters and Elsevier offer CERIF-compatible products.

Management Information

Since CRIS were designed from the outset for managing research information naturally this is their strongest aspect. CERIF-CRIS have evolved (maintaining backward-

compatibility) with the increasingly complex demands of governments, funders, research institutions researchers and others for management information.

Development

We have two major streams of development in recording and managing research information: repositories and CRIS. The former concentrate on research publications (with some limited associated information on persons, organisations) and the latter initially on research projects (with limited associated information on other entities) but in the last decade allowing the point of access to the information to be any research entity (person, organisation, project, funding, publication….).

To meet the evolving requirements of the end-users developments have been attempted, are ongoing and planned.

Repositories extended

Against the background of the REF (Research Excellence Framework, replacing the earlier RAE (Research Assessment Exercise) in UK a JISC (Joint Information Systems Committee of HEFCE (Higher Education Funding Council England) funded a project – Readiness4REF (R4R) - to extend repository systems to encompass the requirements of REF. The repository system used was ePrints based at University of Southampton. While the project demonstrated technical feasibility (in IT one can do almost anything given sufficient resources) the effort required, the difficulty of use and the limitations of the data model adopted led to the conclusion that it is better to use a CERIF-CRIS.

CRIS evolving

As indicated above CRIS generally are evolving to adopt CERIF and the move from CERIF91 to CERIF2000 and beyond has enhanced greatly the applicability of CRIS to the evolving requirements. It is widely accepted that the evolving CERIF standard provides the basis for CRIS meeting the new requirements emerging and expected.

Integration

Repositories exist, are maintained and used, and have – to varying degrees – captured research output in the form of publications. Similarly CRIS exist and are used, capture all aspects of research information with a formal datamodel (CERIF), and are evolving with changing requirements. CERIF-CRIS are capable of representing everything that is stored in repositories whereas repositories are not capable of representing everything stored in CRIS.

This begs the question – why have repositories? Indeed, some research institutions have abandoned their repositories and are using a CERIF-CRIS alone for all functions including those previously done by repositories. However, most research institutions are attempting to integrate their repositories and CRIS. In so doing the major question is ‘what

metadata standard is used and where is the metadata stored'. The starting position is Repositories with DC and CRIS with CERIF. Then the options are:

1. Input DC to repository and CERIF to CRIS;
2. Input DC to repository and copy/convert to CERIF for CRIS;
3. Input CERIF to CRIS and copy/convert to DC for repository;

Clearly (1) is wasteful of resources and error-prone (there are likely to be inconsistencies in metadata in each system) yet in fact is the currently most-practised method, mainly due to organisational rigidity and the ownership of the repository being with the library and the CRIS with the research administration. (2) is impractical since CERIF is a much richer datamodel than DC and so would be essentially empty with only a few attributes (elements in DC) filled. This option is not used. (3) is widely used and especially when the organisational problems between the library and the research administration have been resolved. CERIF to DC (in various instantiations) convertors exist so that – if required – the DC-compatible metadata elements in CERIF can be provided to the repository for local catalog purposes.

An example of (3) is found in Norway. The national CRIS is CRIStin and is (largely) CERIF-compatible. Different research institutions have different repository systems. CRIStin stores metadata for research publications (through a process involving download from Thomson-Reuters Web of Science, clean-up and validation of claiming of publications by researchers) and current work is linking these metadata records in CRIStin to the full text / multimedia articles in the institutional repositories via pointers (URLs). Thus the user has the advantage of the full context of the research (via CRIStin) and also access to the stored scholarly output in the repository.

Conclusion

We expect over the next years further developments. In general, repositories will store the scholarly output full text/multimedia and CERIF-CRIS will store the data/metadata providing the full context of the research. Interoperation will be via CERIF – even for access to publications in other repositories. This is because increasingly the end-user wants not just the research publication but also its context – the project, persons, institutions, funding, facilities/equipment and (very importantly) associated datasets and software. With time the full text/hyperlinked multimedia of scholarly publications will reside either in a simple file store (for fast access, pointed to from the CERIF-CRIS) or within the CERIF-CRIS itself.

Then, the CERIF-CRIS can be used to manage the full research lifecycle of all kinds of users from managers to researchers. For the former research management is evidence-based and justifiable; for the latter the e-Research environment will have arrived (Jeffery 2012).

References

1. CERIF http://www.eurocris.org/Index.php?page=CERIFreleases&t=1
2. CRIStin www.cristin.no
3. DC http://dublincore.org/documents/dces/
4. (Finch 2012) Finch report http://www.researchinfonet.org/publish/finch/
5. (Harnad, Swan 2009) Stevan Harnad, Les Carr, Alma Swan, Arthur Sale and Hélène Bosc: 'Open Access Repositories.Maximizing and Measuring Research Impact through University and Research-Funder Open-Access Self-Archiving Mandates' Forschungsinformation VII, 2009
6. (Jeffery 2012) Keith Jeffery : 'CRIS in 2020' Invited Keynote in Proceedings 11th International Conference on Current Research Information Systems (CRIS2012), Prague June 2012; ISBN 978-80-86742-33-5
7. (Jeffey 2006) Keith G Jeffery and Anne Asserson: 'Supporting the Research process with a CRIS' In: 'Enabling Interaction and Quality: Beyond the Hanseatic League. Proceedings 8th International Conference on Current Research Information Systems (CRIS2006), Bergen 2012; ISBN 978-90-5867-536-1
8. MARC http://www.loc.gov/marc/
9. OAI-PMH http://www.openarchives.org/pmh/
10. ORCID http://orcid.org/
11. Readiness4REF (R4R) http://www.jisc.ac.uk/whatwedo/programmes/inf11/sue2/r4r

Repositories, CRIS and CRIStin: the story so far – prezentácia z konferencie

Repositories, CRIS and CRIStin: the Story so Far

Anne Asserson
Digital Systems and Services
University Library

UNIVERSITY OF BERGEN

- **Repositories**
- Structured systems and Managment information
- CRIStin
- Open Access
- Integration

2 major streams of Development in recording and managing scholar information

- Repositories
- CRIS (Current Research Information System)

- Repositories store objects (full text, multimedia, and datasets)
- Keep ca 15% of the institutions full text
- Metadata recording the existence of articles, books and chapter i books will of course be higher.
- Different attitude to what is stored
  - Only Peer reviewed
  - All research output

The low degree of 'fill' of repositories can be ascribed to several factors:

(a) the threshold barrier to input the metadata and load the article;

(b) pressure and restrictions from commercial publishers through copyright and other legalistic mechanisms;

(c) lack of perceived benefit.

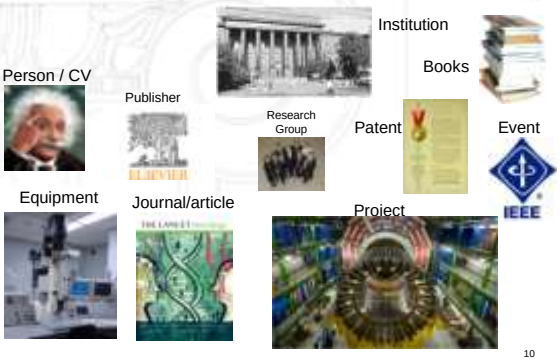
- OAI-PMH is used as the interoperation protocol allowing harvesting of metadata from any compliant repository.
- The syntax of DC elements from different repositories can be combined together as long as the simple element structure is followed;
- the use of repeating groups or extended namespaces frustrates interoperation.
- Multilinguality and differing semantics (meaning of the lexical string) makes interoperation without much human intervention impossible.

- Repositories
- **Structured systems and Managment information**
- CRIStin
- Open Access
- Integration

- Increasingly there is a requirement –from research funders and research institutions to relate scholarly articles to research funding (grant or grants) as well as to person(s) and organisation(s).
- and when appropriate - to relate the publication to particular research facilities and/or equipment.

- The purpose of a CRIS (Current Research Management System) is to record research information for the purposes of management and utilisation by a
  - funder
  - research institution or researcher and
  - for utilisation by innovators, entrepreneurs,
  - the media and the public.

…..typical Research Objects in a CRIS



CRIS Requirement at Institution level

- a tool for policy making
- evaluation of research based on outputs
- document the research activities
- document research output
- a formal log of research in progress
- to assist project planning.

….for Reseachers /end users

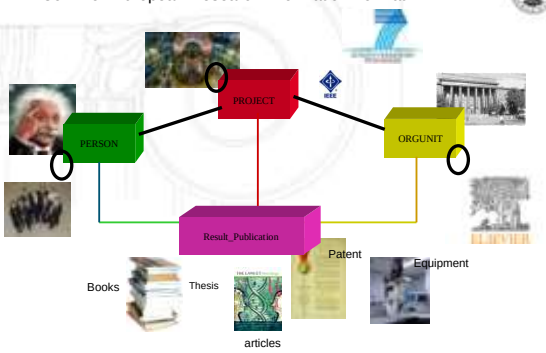
- avoid duplication of research activity
- to evaluate opportunities for research funding
- analyse research trends, locally, regionally and internationally
- references/links to full text
- locate new contacts/networks
- identify new markets for products of research

CERIF short history

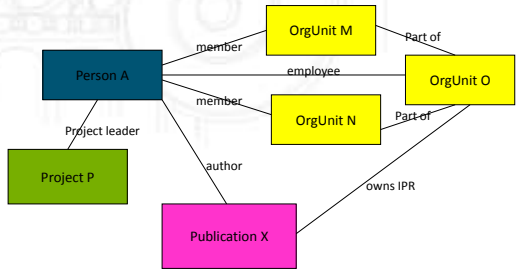
- 1987-90 European experts to define a Common European Research Information Format CERIF
- 1991 a single record format and a research classification scheme (not updated from 1988)
- Research documentation and CRIS more important
- 1997 ERGO European Research Gateways Online) pilot – a single catalogue of projects from national databases, 20 countries 90.000 records
- 2000 CERIF model

- It is clear that repositories using DC (Dublin Core) for metadata are incapable of meeting these new requirements for management information on research.
- Such as
  - benchmarking institutional performance and for
  - research funders to assess the research performance against funding for researchers and institutional organisations

CERIF Common European Research Information Format

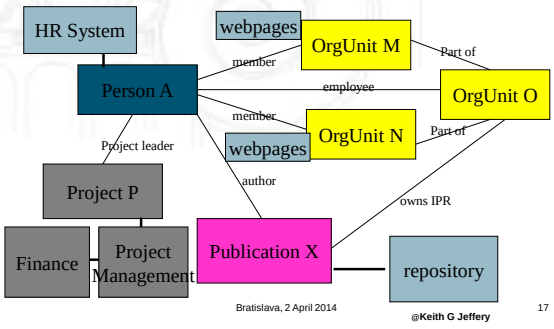


CERIF (Current Research Information Format) Model

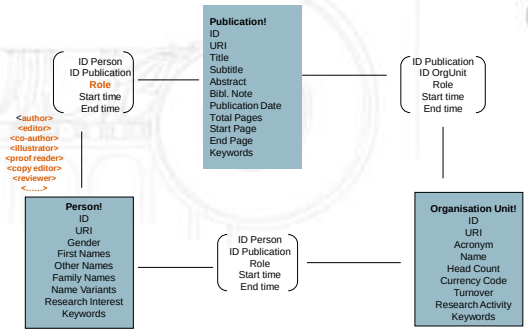


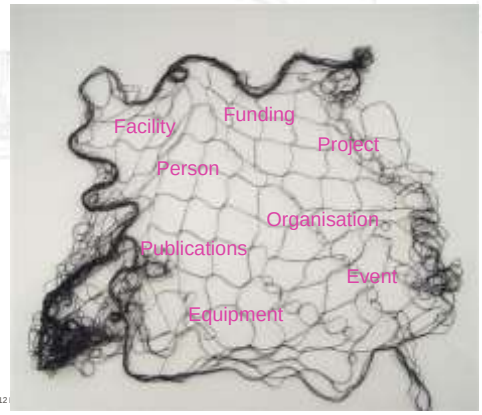
@Keith G Jeffery

CERIF Model



@Keith G Jeffery





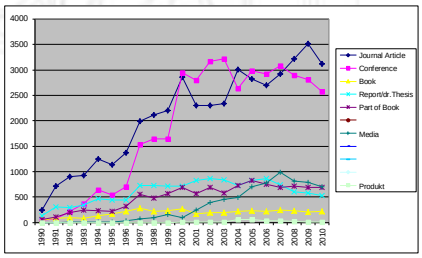
- Repository
- Structured systems and Managment information
- **CRIStin**
- Open Access
- Integration

Increasingly CRIS are becoming CERIF-compatible
(a) to permit interoperation but more importantly
(b) because the model reflects accurately the real world of research information.

Several commercial companies offer CERIF-compatible CRIS and both Thomson-Reuters and Elsevier offer CERIF-compatible products.

CRIStin national CRIS in Norway

UiB output 1990-2010

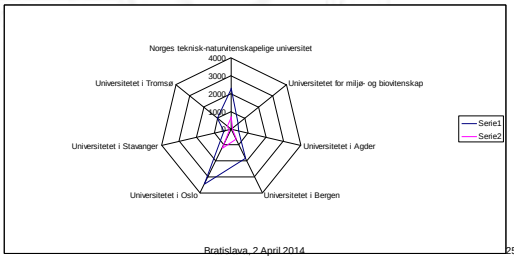


Researchers profile

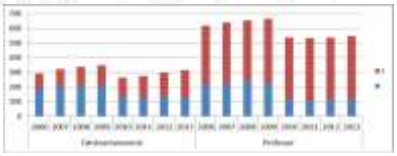
Fulltext via DOI

Level 1 (80 % of publication channels)
Level 2(20 % of publication channels)

Norway populate with ISI Thompson Reuters data
No assessing or evaluation related to citations or impact factor



Who are publishing?

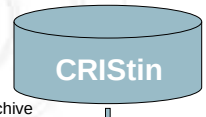


- Repository
- Structured systems and Managment information
- CRIStin
- **Open Access at UiB**
- Integration

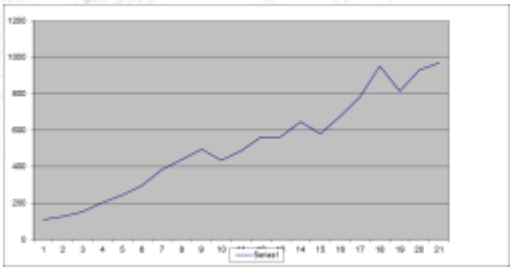
CRIStin – metadata to fulltext

CRIStin holds the contextual metadata to the Institutional Repositories of the research performing institutions

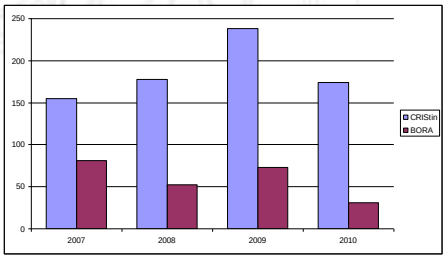
- •BORA Bergen Open Research Archive
- •DUO
- •MUNIN
- •Hera
- •…….



Dr. Thesis 1990-2010



Dr.thesis



- 2013 The University Board at UIB agreed to establish fund and run a pilot for 3 years to support Open Access Publishing
- 1,5 mill NOK (200 000 euro)was allocated for publishing Gold Open Access
- 2014 the fund has increased to 1,8 millioner NOK
- Researchers at UiB can apply to cover expenses for GOLD pubishing

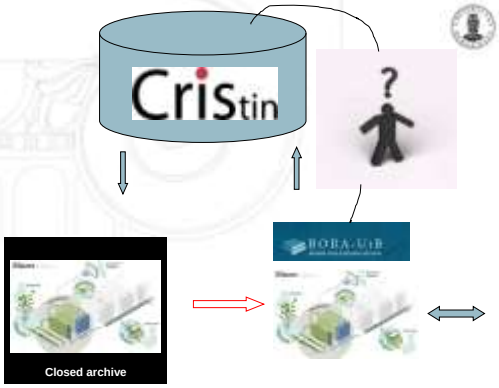
- Faculties are allocated a given sum
- The publications has to be deposited in the local repository, BORA
- Today this has lead to 80 articles, but expect more
- OA in Hybrid journals, not necessarily OA
- First year is now under evaluation

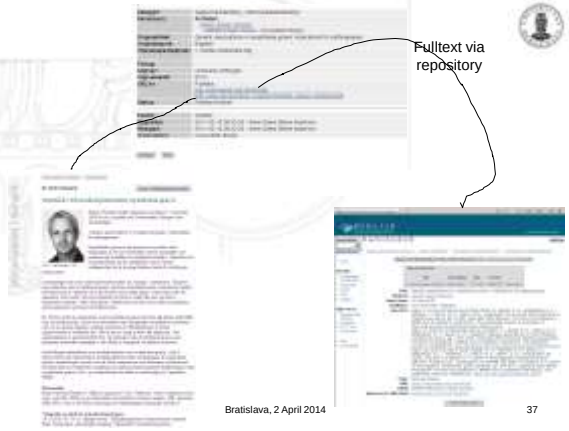
- Repositories
- Structured systems and Managment information
- CRIStin
- Open Access at UiB
- **Integration**

CRIStin as metadata to fulltext

2 ways to reach the full text

- Copy of fulltext in the Institutional Repository
- DOI (Digital Object Identifier) links to the article, but dependent on subscription.





- Repositories exist, are maintained and used, and have captured research output in the form of publications.
- Similarly CRIS exist and are used, capture all aspects of research information with a formal datamodel (CERIF), and are evolving with changing requirements.
- CERIF-CRIS are capable of representing everything that is stored in repositories whereas repositories are not capable of representing everything stored in CRIS.

Why have repositories?

- Some research institutions have abandoned their repositories and are using a CERIF-CRIS alone for all functions including those previously done by repositories.
- Most research institutions are attempting to integrate their repositories and CRIS.
- In so doing the major question is ‘what metadata standard is used and where is the metadata stored’.

The options are:

Input DC to repository and CERIF to CRIS

This is the currently most-practised method, mainly due to organisational rigidity and the ownership of the repository being with the library and the CRIS with the research administration.

Input DC to repository and copy/convert to CERIF for CRIS;

This impractical since CERIF is a much richer data model than DC and so would be essentially empty with only a few attributes (elements in DC) filled.

Input CERIF to CRIS and copy/convert to DC for repository

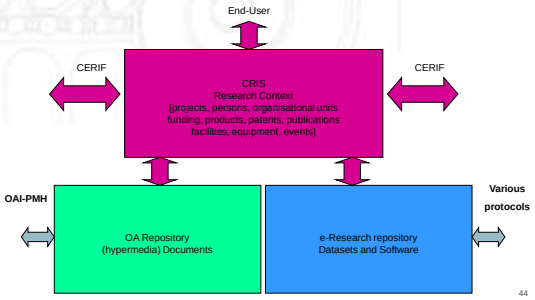
This is widely used and especially when the organisational problems between the library and the research administration have been resolved.

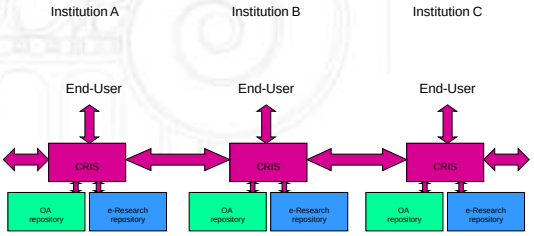
CERIF to DC (in various instantiations) convertors exist so that – if required – the DC-compatible metadata elements in CERIF can be provided to the repository for local catalog purposes.

We expect over the next years further developments.

- In general, repositories will store the scholarly output full text/multimedia and
- CERIF-CRIS will store the data/metadata providing the full context of the research.
- Interoperation will be via CERIF – even for access to publications in other repositories.
- Increasingly the end-user wants not just the research publication but also its context – the project, persons, institutions, funding, facilities/equipment and (very importantly) associated datasets and software.

CERIF-CRIS and the Repository are an essential part of a e-infrastruture





CRIS 2020

……Then, the CERIF-CRIS can be used to manage the full research lifecycle of all kinds of users from managers to researchers.

For theresearch management is evidence-based and justifiable; for the researcher the e-Research environment will have arrived….

'CRIS in 2020' Keith Jeffery , Proceedings 11th International Conference on Current Research Information Systems (CRIS2012)

UNIVERSITY OF BERGEN
University Library