euroCRIS
Current Research Information Systems

CRIS SEMINAR SEPTEMBER 2009
HOTEL ERRERA, BRUSSELS

Recording Research

Report September 2009

euroCRIS acknowledges gratefully the strategic partners who co-developed the programme, provided speakers and other forms of support.

Hospitality was provided by the Flemish Ministry Science Division for which we are extremely grateful. euroCRIS acknowledges also the kind provision of the venue, Hotel Errera, the Official Residence and Reception Building of the Flemish Government

Ministry of the Flemish Community
Science and Innovation Administration

The organisations provided (along with euroCRIS) speakers, other attendees, funding and other assistance which has made the seminar possible.

The seminar partners agreed that a report should be produced by euroCRIS, agreed with the partners, published for all the partners and placed on the euroCRIS website together with the presentations.

www.eurocris.org

euroCRIS it the not-for-profit organisation promoting best practice in CRISs, spanning the field from raw experimental and simulated data through research management systems to research publications. Through its biennial CRIS conference and annual strategic seminars euroCRIS disseminates best practice. euroCRIS members' meetings share experience and set policy. EuroCRIS maintains CERIF, the Common European Research Information Format.

1. EXECUTIVE SUMMARY

Following previous successful seminars covering: CRISs (Current Research Information Systems); Open Access to Research Information; the Research Information System and the Research Process in a GRIDs Environment; the Relationship between Innovation and CRISs; CRIS and the e-infrastructure and CRIS at the Centre of the Research Business, euroCRIS identified for this 2009 seminar the need for a focus on Recording Research. The resulting conclusions are as follows:

1. recording research is critically important: research is expensive and the outputs should be (a) recorded (b) openly available for use;

2. to achieve this research funding and research performing organizations need a CRIS (Current Research Information System);

3. the recording in the CRIS should be in a form that allows interoperation and homogeneous access over heterogeneous sources: CERIF is the current EU recommendation to member states for this purpose;

4. the threshold barrier to (meta)data collection has to be overcome by (a) using incremental steps with re-use of already collected data (b) intelligent user interfaces to assist input;

5. increasingly the move to evaluating research at national level drives the need for a CERIF-CRIS to reduce the effort needed to produce the required information;

6. with a CERIF-CRIS it is also possible for a research organization (funder or performer) to utilise the management information for strategic decision-making.

2. PURPOSE

The background is that both nationally and internationally there is much information from research activities, but it is fragmented, varies in quality and structure and is not easy to locate, retrieve and use. Furthermore commonly it is not curated for use by others who are not familiar with the data. euroCRIS identified the need for policies and processes to assure the recording of research such that it has appropriate quality, persistence, availability and is described suitably for re-use.

The research information is then used for several purposes:

1. as management information for decision-making in the research organization;

2. to report research outputs to evaluation systems nationally;

3. to allow researchers faster/cheaper access to information about research ongoing in their own organsation and more widely

4. to expose the intellectual property and achievements of a research organization to all stakeholders;

5. to encourage knowledge and technology transfer from the research organization to industry leading to wealth creation;

6. to make available the results of research as a feed into teaching and learning;

7. to make available information on research - commonly via the media – to the general public

For these uses of the information to be valid and reliable it is essential that the information has appropriate quality and is curated so that its use is subject to understanding of context, achieved by appropriate metadata use.

3. TAKE HOME MESSAGES

The key messages from the partners in the conclusion of the seminar were as follows:

1. CERIF-CRISs are a key component of e-infrastructure, an institutional asset, necessary for research strategy and management and are needed – and need to interoperate - across research funding and performing institutions;

2. CERIF-CRISs encourage evaluation, research and researcher improvement, knowledge/technology transfer, innovation for all stakeholders and especially assist the role of research manager;

3. CERIF-CRISs make provision of material for evaluation (internal or external) much more effective and efficient;

4. TOPICS COVERED IN THE PROGRAMME

Session 1: - Why Recording Research
- The purpose of recording research and the issues we are meeting, *Keith Jeffery, STFC Science and Technology Facilities Council*
- Presentation of "Readiness for REF" (or "R4R") - to help universities meet the requirements of the Research Excellence Framework, *Stephen Grace, King's College, University of London*
- Improving the impact of research output from a university, *Les Carr, University of Southampton*
- Managing research information as an institutional asset, *Anna Clements, University of St. Andrews*
- The value of recording each step of the research process, *Keith Jeffery, STFC Science and Technology Facilities Council*

Session 2: What is/should be recorded?
- The research lifecycle end-to-end, *Malcolm Read JISC Joint Information Systems Committee*
- Identifying persons unambiguously: the DAI (Digital Author Idenatifier) project in The Netherlands, *Chris Baars, KNAW Royal Netherlands Academy of Arts and Sciences*
- Metadata for all research information – CERIF, *Brigitte Joerg, DFKI , German Research Centre for Artificial Intelligence*

Session 3: Who does/should record it?
- The role of the researcher, *Jüri Engelbrecht, ALLEA All European Academies*
- The role of the research manager, *Martin Jägerhorn, AVEDAS AG*
- The role of the librarian / curator, *Maria Heijne, Delft University of Technology Library*

Session 4: How should research be recorded?
- Manual (meta)data collection – problems and solutions, *Stijn Christianens, Collibra*
- Automated (meta)data collection – problems and solutions, *Tina Lingjærde & Andora Sjøgren, USIT – Norwegian Center for Information Technology Services*
- Timeliness and quality, *Jürgen Güdler, DFG German Research Foundation*

Session 5: Once recorded, how is it used, by whom, and for what purposes? How is it protected? How is it curated/preserved for future generations?
- Use by innovators, *Nikos Houssos, EKT – Greek National Documentation Centre*
- Use by a researcher, *Wouter Spek, APA – Alliance for Permanent Access*
- Use by a research manager, *Dag Aksnes, NIFU – Norwegian Institute for Studies in Innovation, Research and Education*

Session 6: Discussion and Take home messages
- Scientific data information domain of e-Infrastructures. Activities and plans, *Carlos Morais-Pires, European Commission, DG INFSO*

# 5. KEY DISCUSSION POINTS FROM THE SESSIONS

WHY RECORDING RESEARCH

The purpose of recording research and the issues we are meeting

We record in order to inform others, to create strategy, to stimulate further research and innovation. Also there is an expectation of a complete record describing the research, relating all the entities by role and by time to give a reconstruction of state at any time. We record information on core entities and their inter-relationships (roles) through time.

Key questions concern: What research was done (project); Who did it (person, organisational unit); Who funded it (organisational unit); What was produced (products, patents, publications; note: products include data and software), How was it done (facilities, equipment); When was it disseminated (events). We need to record temporal information and roles between entities.
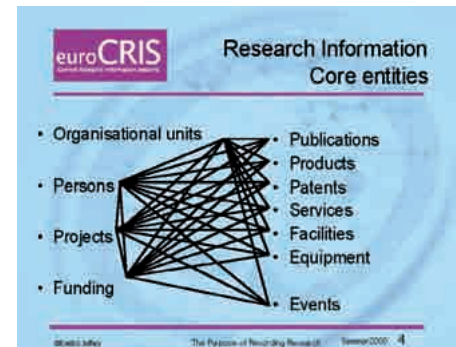
Readiness for REF" (or "R4R") to help universities meet the requirements of the Research Excellence Framework

In UK the Research Excellence Framework is the successor evaluation mechanism to RAE – the Research Assessment Exercise. The purpose is external peer assessment of the quality of research in a given subject area. REF/RAE informs research funding from the government associated with education; this is separate from competitive research proposals funded by the research councils. The assessment criteria for REF are not yet published but it is expected that there will be more metrics that are generated by computer hence the need for CRIS and institutional repositories of publications. The R4R project decided to apply CERIF to REF generating a CERIF4REF application profile, utilized by CERIF plug-ins for ePrints, DSpace, Fedora. The project has a high profile in UK universities.


R4R project

- Survey existing university systems
- Case studies in HEIs , funders, publishers
- Analyse REF requirements
- Apply CERIF to REF
- CERIF4REF application profile
- CERIF plug -ins for ePrints, DSpace, Fedora
- Demonstrators at King's and Southampton

Improving the impact of research output from a university

There is an Open Access Advantage: EA + QA + UA + (CA) + (QB) EA: Early Advantage: Self-archiving preprints before publication hastens and increases usage and citations (higher quality articles benefit more: top 20% of articles receive 80% of citations) QA: Quality Advantage: Self-archiving post prints immediately upon publication hastens and increases usage and citations (higher-quality articles benefit more) UA: Usage Advantage: Self-archiving increases downloads (higher-quality articles benefit more (CA: Competitive Advantage): OA/non-OA advantage (CA disappears at 100%OA, but very important today!) (QB: Quality Bias): Higher-quality articles are self-selectively self-archived more (QB disappears at 100%OA)
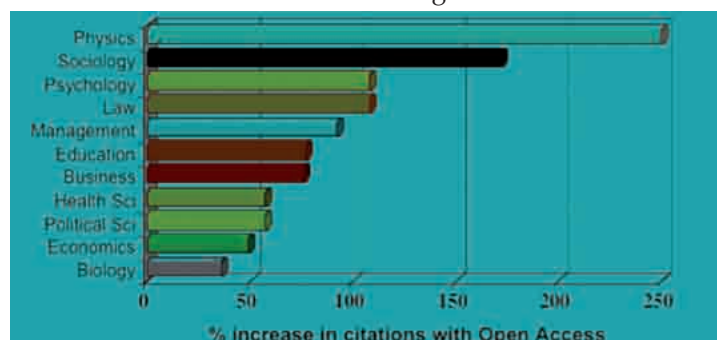
## Managing research information as an institutional asset
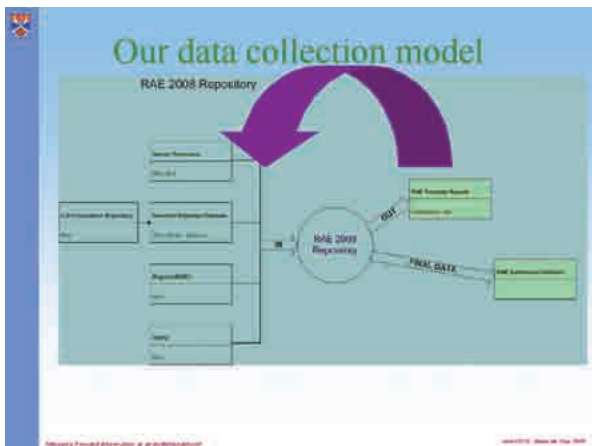
RAE/REF determines a third of research income in UK > £1.5 billion. The challenge for any university of excellence is to bring data from across the university together. RAE2008 used legacy systems and procedures that were not automated enough and the policies were not clearly defined and enforced. Since then there has been a review of business processes and some systems replacement. Data is an asset: it should be managed as such. Key aspects are: Collect what is required: know our customers needs. Assign Trustees: responsible for quality and relevance. Share data: capture once as close to the point of creation as possible then make it accessible and understandable. Use standards [CERIF]; use authority sources. Comply with data protection and security policies.

The value of recording each step of the research process

Numbers of researchers and output per researcher (Publications, Patents, Products, metadata for research datasets) are increasing. It is an effort to catalog i.e. input metadata. The load is too great (for the user) and does not scale (with increasing numbers). The research process is a workflow with e-forms. At each step (meta) information is required and stored incrementally (so assuring re-use, minimal effort). The researcher sees benefit from the process: examples include automated CV; automated publication list. With a CRIS the researcher can track competing and cooperating teams while her research becomes visible to intermediaries for exploitation. A CRIS can

provide boilerplate information for research proposals. CERIF records these common data elements and is extensible through the linking relation mechanism Relationships between entity instances (e.g. documents) can also be expressed explicitly (i.e. asserted) e.g. references and / or citations can be recorded by directly inputting the information into the CERIF-CRIS.
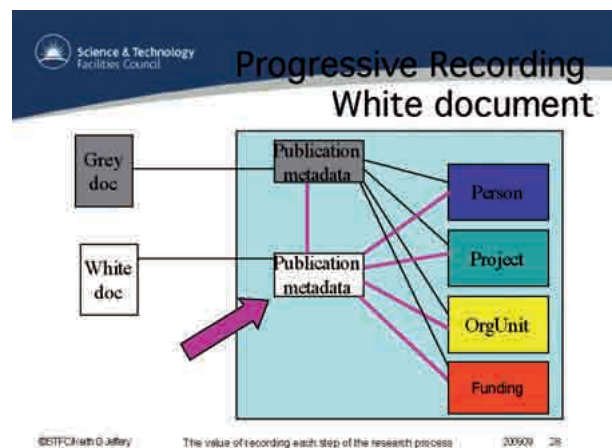
WHAT IS/SHOULD BE RECORDED?
The research lifecycle end-to-end

There are implications for CRIS because Research funders and universities are increasingly called on to justify their work as public spending is reduced to address huge government debt. It is very difficult to measure RoI (Return on Investment), because of issues of attribution, traceability, measurability, time lags. There is a need to blue skies research – concentrating too much on research with immediate/obvious impact is a short term policy, and detrimental long term. Opening up universities' knowledge assets and expertise to enable external parties to consult, enrich and co-develop those assets creating new knowledge, new markets and new innovative opportunities benefits the economy and society. Enabling collaboration between academe and industry gives technology based businesses significant competitive advantage where content, specialised analysis, presentation, visualisation and design are critical to competitiveness.


4

Identifying persons unambiguously: the DAI (Digital Author Identifier) project in The Netherlands

There is a need to avoid name ambiguity and variants to keep track of people across institutes during their academic career and to distinguish people with the same name. In 2004 a national approach and national thesaurus of author names was initiated. DAI helps to connect information with other systems. All researchers of universities & KNAW have a DAI (45.000). The thesaurus is hosted by OCLC. Increasingly DAIs are used in open access publications. Plans are made for non-university institutes to be included. There is a need for synchronisation between CRIS and IRs. The DAI is not only about 'giving every person a number', but also about authenticity of scientific output.
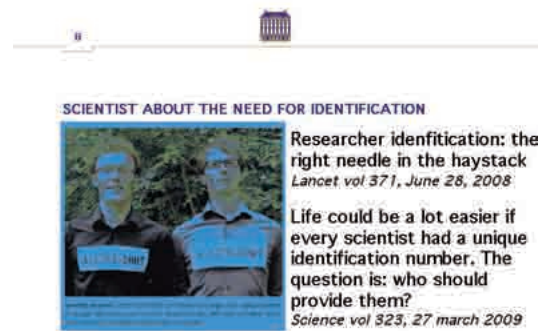

Metadata for all research information – CERIF (COMMON EUROPEAN RESEARCH INFORMATION FORMAT)

CERIF deals (Conceptual level) with Research Entities and their Relationships. From this specification a Model (Logical Level) provides the basis for the implementation (Physical Level).
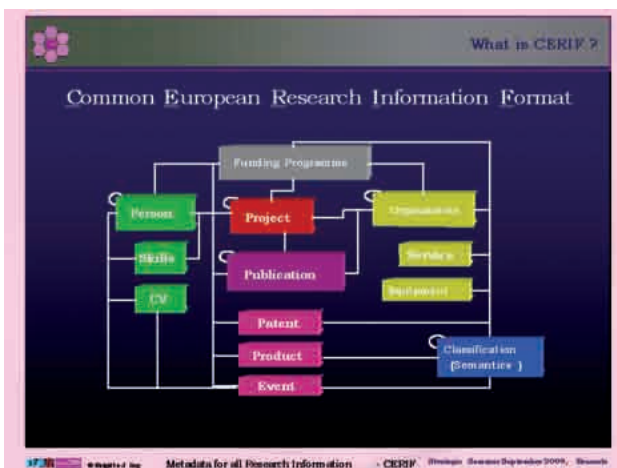
Concepts include: "A research a project is mostly funded by a Funding Agency." "In research, journal impact is measured by citation counts." "A researcher produces knowledge in publications." "A University is an organisational unit that has departments doing Research that produce Results."

Typical questions might be: How many projects has the EC funded in FP6? How many articles has author X published in 2007 as a first author? How often have articles by author X been cited? Did author X publish with institutionally external authors? In how many FP7 projects does organisation Z participate? How many publications have resulted from project Y? How many people have been employed in the course of FP6 projects from the 1st call in the NMS? How many PhD students have participated in FP6 projects? How many women have been involved in FP6 projects? How often have articles in journal A been requested in 2007? How many articles have been published in the field of B?
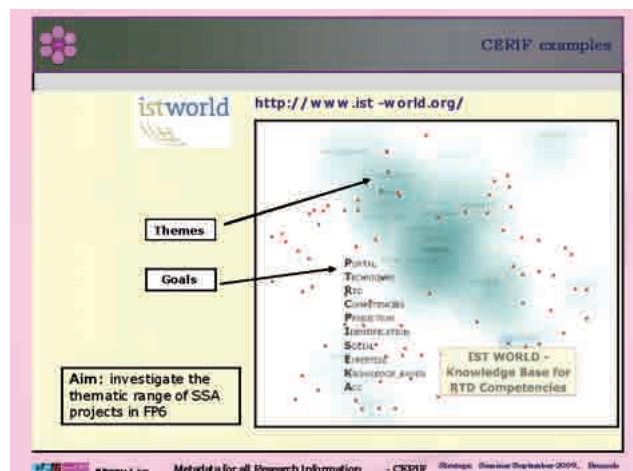
WHO DOES/SHOULD RECORD IT?

The role of the researcher

The role of the researcher includes recording and presenting her research as well as performing the research. (Benjamin Franklin: to study, to finish, to publish) The new knowledge is within a context of science, society, information, applications, education and has research ethics and research integrity. Recording is the basis for new ideas, the basis for applications and advice for society. In the future there will be text mining and automatic processing. Meantime we need to

address principles of scientific integrity, dealing with allegations of misconduct, recommended rules for procedures. Researchers should endeavour to publish as early as possible in an open, honest, transparent and accurate manner. Commercial considerations or patent application may require some delay. Authorship must be based on a creative contribution to the research, honorary or 'ghost' authorship is unacceptable. Authors are responsible for the content of the publication including acknowledgements and citations.


The role of the research manager

The research manager considers the management of research from a perspective of what is best for the organization and its researchers including finance. She takes into account other organizations for competition or collaboration. For this she needs a CERIF-CRIS (ideally linked to those of other research institutions). A research manager plays many different roles at the same time, closely interacting with the researchers and other research managers in the workflows. The workflows have users in multiple roles at multiple entry points. This has implications for a CRIS but is matched by the evolving CERIF specification. Key advice is a) to avoid spending resource on collecting data that won't be needed - reuse & quality assure. b) Integrate with external and internal systems think CERIF! c) input once, output in many ways ensuring effectiveness and efficiency.

The role of the librarian / curator

A Project (2008-2010) financed by 3 Technical Universities (Delft, Twente, Eindhoven) is underway to ensure long term digital data preservation in the technical/science environment. It is based on successful (SURF) projects in Delft: e-Archive and DARELUX. The project involves the exploitation of a 'data-archive' to facilitate preservation of 'static' data of data producers and

advises data producers how to preserve 'dynamic' data in a collaboratory environment for access and publication ('data-lab'), e-science can/will lead to different methods of knowledge exchange through publication, visualization, data exchange and involving access to large data collections/large scale computing resources. The core business is information/data management (storage, preservation, metadata creation, access reuse of data/information). Librarians will become data consultants, data service providers, data analysts, data miners, data curators.
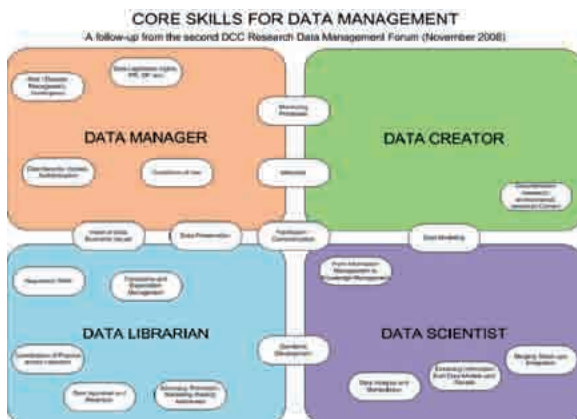
### HOW SHOULD RESEARCH BE RECORDED?
Manual (meta)data collection – problems and solutions

The problem lies in repeatedly interpreting the meaning of different formats: no reuse; no documentation, trial & error, repeating errors, no knowledge sharing. This is not manageable, not scalable and just increases the legacy with attendant high cost.

From a series of statements at the conceptual level, and using the tool illustrated above, it is possible to derive Basic facts:

FundingProgramme has FundingProgrammeId; FundingProgramme has Valid Period; Valid Period is between StartDate and EndDate; FundingProgramme has Budget expressed in known Currency; FundingProgramme has Name expressed in known Language and Translation; FundingProgramme has Description expressed known Language and Translation; FundingProgramme has Keywords expressed in known Language and Translation; FundingProgramme has Classification; FundingProgramme has URI; FundingProgramme is part of FundingProgramme;
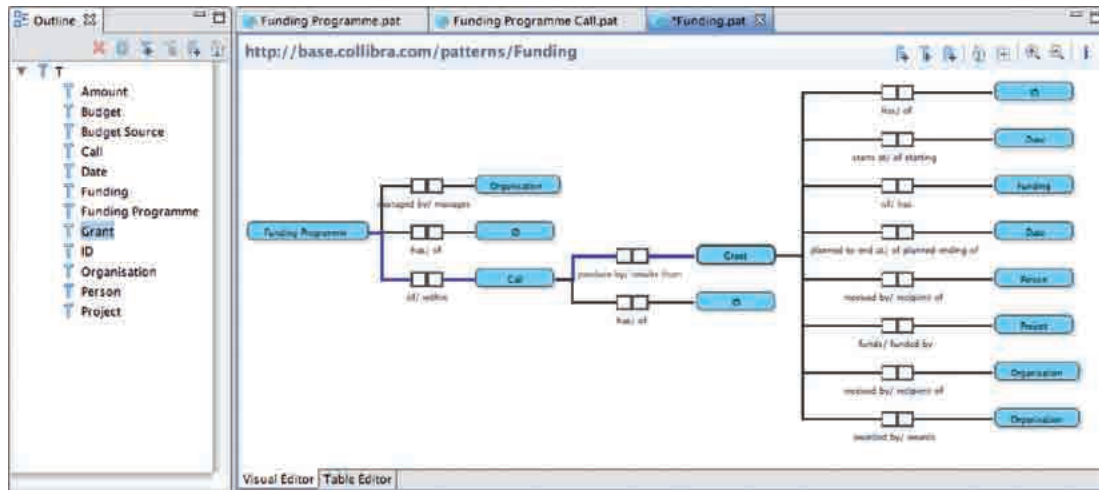
The project has adopted a semantic tool with CERIF output. It provides a generic base model with agreed semantics to stakeholders

Automated (meta)data collection – problems and solutions

FRIDA is an integrated research environment for the documentation and presentation of research activities, research results and scientific competence. It generates statistics for research activities at Norwegian universities. FRIDA plays a major role in the annual funding of universities by the Norwegian Ministry of Education and Research. Automated metadata collection is as follows: there is direct import from ISI and Norart. The system contains registers/separate tables of: periodicals, series, publishers, organizations (institutions), common code tables. The common use and maintenance of these registers is an important quality measure. For each user in FRIDA, a personal record must be imported from the institution's user administrative system (again reducing effort). FRIDA can automatically update the code for an organisation unit to another code.
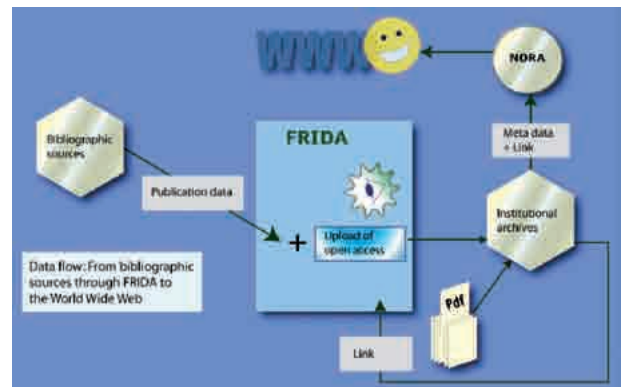

Timeliness and quality

DFG produces Information products based on process-produced data: Information about proposals (approx. 450,000); Information about institutions (approx. 30,000); Information about persons (approx. 150,000). In addition there is information about programmes, regions, countries, disciplines.

The Data in ElektrA are very up-to-date because funding-process related products (e.g. letters to applicants and reviewers) provide the information. Quality responsibility is decentralised Nevertheless there is a need for a central quality assurance unit for the development of complex quality assurance routines. Outputs include statistical reporting for the statutory bodies of the DFG. Good quality data is costly and is labour-intensive.
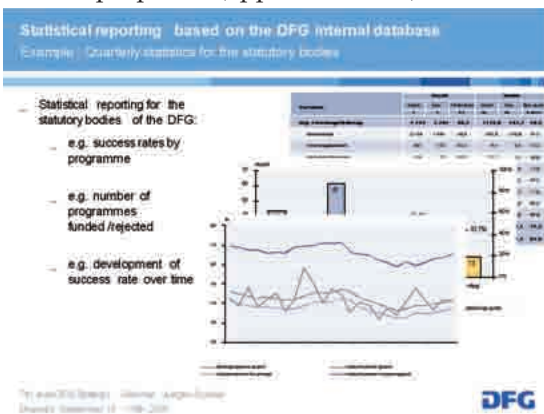
ONCE RECORDED, HOW IS IT USED, BY WHOM, AND FOR WHAT PURPOSES? HOW IS IT PROTECTED? HOW IS IT CURATED/PERSERVED FOR FUTURE GENERATIONS?
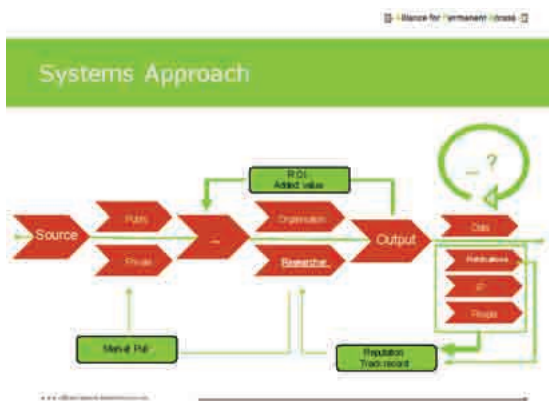
Use by innovators

Enterprise Europe Network (EEN) is a strategic alliance of the most important SME support networks (EICs & IRCs). EEN provides a One–Stop-Shop for Information, feedback and both international business cooperation services and Innovation, Technology & Knowledge transfer services. EEN brings together academia and industry and also industry with venture capitalists thus stimulating innovation. The innovation may be via a spin-out company. There is a real need for databases of research information including patents and products.

Use by a researcher

Once recorded there are questions to be addressed: How is it used? By whom? For what purposes? How is it protected? How is it curated/preserved? The Alliance for Permanent Access to the Records of Science PARSE. Insight project is developing a roadmap and recommendations for building the e-Science infrastructure in order to maintain long-term access and use of scientific digital information in Europe. New knowledge and know-how depend more and more on the ability to re-use and share data. As well as a technical dimension there is a policy dimension. The OECD Principles and Guidelines for Access to research Data from Public funding 2007 is one example. There is also a business model dimension (what is the value of data and who will pay) and a legal dimension encompassing the ownership issue and the IP-paradigm.

Use by a research manager

NIFU STEP is a Norwegian research institute doing contract research, analyses and consultancy for various organisations such as ministries and research councils. Data on all types of scholarly publications in all fields of research in the Higher Education Sector are collected through two documentation systems (FRIDA and ForskDOK). Citations are used as an indicator scientific impact and international visibility. Co-authorship used as an indicator of collaboration. Traditionally, bibliometric analyses of research performance have usually been based on the ISI-database (Web of Science, Thomson Reuters). In Norway FRIDA (CERIF-CRIS) is well suited for bibliometrics: the bibliographic references in the documentation system are standardized and analyzable by publication channel and type of publication; it has complete, verifiable and structured data; it covers all scientific/ scholarly publications not only journal articles and non-scientific/non-scholarly publications can be excluded. The funding of research institutions is now partially based on these bibliometrics.

Figure 1. Publication types* in five major areas (N=24,154)
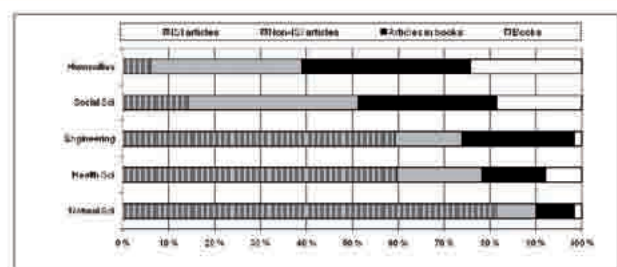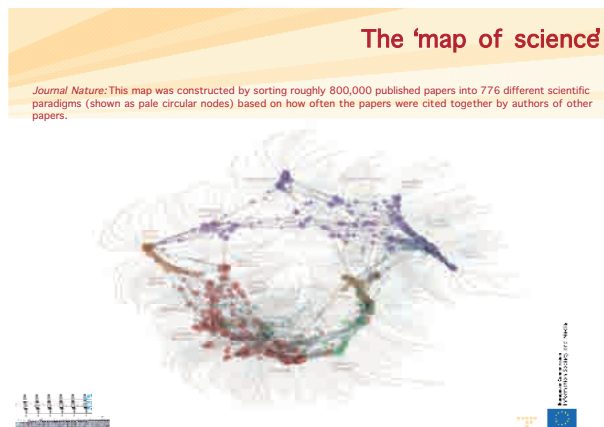
*)Weighted: 1 book = 5 articles

The researchers do not have to submit their publication lists anymore In addition to using FRIDA as the main data source, information on the "level" of the journals and book publishers are used in the analyses.

Scientific data information domain of e-Infrastructures. Activities and plans

A Communication to the Council and European Parliament states that e-Science and e-Infrastructures are to be priorities in Europe. The EC is actively participating and influencing global partnerships in science. There is coordination with EU Member States to put the strategy in practice. A recent communication to European Institutions highlighted the importance to embrace the e-Science paradigm shift. This includes the strategic role of e-Infrastructures as a crucial asset underpinning European research and innovation policies. The EC is calling for a concerted effort to address science and ICT global scientific challenges with high societal impact. Problem are based on geography and cultural heterogeneity are complexity, volumes of information, quality of information, incentives to share, organisational barriers. e-Infrastructures. encourage collaboration, with processing, simulation, repositories of data and curation/review, Accessibility to research data is important for: the good management of public investment, creation of strong value chains of innovation, enhancement of value from international co-operation. The emergence of "big data science" is long-term and global. It reflects the increasing value of observational and experimental data in virtually all fields of science. Multi-disciplinary approaches, new participative paradigms and global research communities are an essential part of the strategy… but organisational, governance and financing models need reconsideration, informed by sociological and cultural considerations.

The 'map of science'

Journal Nature: This map was constructed by sorting roughly 800,000 published papers into 776 different scientific paradigms (shown as pale circular nodes) based on how often the papers were cited together by authors of other papers.

6. CONCLUSIONS

1. ALLEA stated that its institutes needed a CRIS, to ensure research integrity, to provide advice based on evidence to justify the science budget. Although at present euroHORCS were sceptical about a global information system we had to make the case, demonstrating why it is necessary.
2. ESF stated that their currently being renovated IT system is using CERIF as their conceptual basis and as their view to the world.
3. EC recognize the importance of CRIS in the e-infrastructure.
4. APA emphasised the need for new infrastructure including curation with CRISs playing a key role.
5. Greynet wish to engage in a study on how CRIS can drive the utilisation of Grey resources.
6. JISC emphasised that open innovation would only be realised through CRIS recording the research.

It was emphasized that the scope of CERIF-CRIS should be clarified to cover all relevant information; the openness should be emphasized including supporting OA IRs (Open Access Institutional Repositories); the interoperability should be promoted and their use as a link between research funders and performers has great potential.

POSTSCRIPT

euroCRIS in a nutshell

A Current Research Information System, commonly known as "CRIS", is any information tool dedicated to provide access to and disseminate research information, such as People, Projects, Organizations, Results (publications, patents and products), Facilities, and Equipment.

EuroCRIS is established to address issues of current research information systems (CRIS) worldwide, but with emphasis on Europe. CRIS may be organized thematically or along the lines of scientific disciplines. Issues are, but not limited to: databases global, thematical and according to type of information (expertise, projects, institution, facilities and products - including publications); standards and guidelines; best practice; data access and exchange mechanisms; and to address other data standardization issues within the realm of research, training and development (RTD, R&D), in a timely and efficient manner.

The primary goals of euroCRIS are to act as a single forum for all interested individuals and organizations to enter into dialog and resolution of all matters related to the use of information technology in the conduct of all research information system business. euroCRIS supports standardized, streamlined information exchange across all aspects of the CRIS lifecycle as follows:

- Promote and improve communication and interaction between global CRIS
- Maintain and publish the CERIF (Common European Research Information Format) recommendation and any standards endorsed by euroCRIS
- Organize and run the CRIS series of conferences with associated workshops and other events
- Provide a source of expertise in CRIS to members and to others under business arrangements made at the time
- Develop euroCRIS guidelines
- Nurture the CRIS community by events, a monthly newsletter, an online discussion forum and other appropriate mechanisms
- Provide a forum for exploring and exploiting new and emerging concepts and technologies (including data quality, standards, etc.)
- Establish a one-stop portal / gateway to international CRIS resources

Membership is open to any organization or individual who is interested in CRIS - without geographical limitations. Structure, identity and procedures have been laid down in formal statutes available on the website. For adequate and flexible functioning the completion of major aims and core actions is assigned to Task Groups, as stipulated in the statutes. At present four Task Groups are established and fully operational.