# Research product repositories: strategies for data and metadata quality control

Luisa De Biagi*, Roberto Puccinelli**, Massimiliano Saccone*, Luciana Trufelli*

*National Research Council of Italy*

*CNR Central Library 'G. Marconi'
**CNR Informative Systems Office

**Abstract**
In recent years a significant effort has been spent by R&D institutions and scientific information stakeholders in general to enhance and improve the quality of Open Access initiatives and the performance of the associated services. Nevertheless much work is still needed to tackle pending data quality issues.

This paper proposes some functional and organizational solutions, based on the cooperation of all the main actors of the R&D system, which in our view should help improving quality control of data and descriptive metadata stored in research product Open Access (OA) repositories. We think that this strategy could favor a substantial innovation of the document management services offered to the scientific community and to policy makers, ensuring the interoperability between institutional repositories and Current Research Information Systems (CRIS).

Particular emphasis is given to the problem of data and metadata indexing and organization with respect to unconventional research products, which represent an important asset in the field of scientific communication.

# Introduction

In Europe, despite the efforts of the scientific community and of many expert groups, effective methods and tools for R&D performance evaluation are still not available. This, in our opinion, is a top-priority issue, since reliable measurements are a pre-condition for credible process and product quality assessment.[1]

In this paper we propose a cooperative organizational approach for tackling some crucial challenges, such as research product metadata quality certification, with particular focus on metadata stored in Open Access repositories (OA).

Currently some national evaluation systems[2] [3] leverage data coming from institutional repositories, which are integrated within R&D Information Systems[4]. Disciplinary and institutional repositories can be used as data sources for R&D performance measurement, also because they keep products which are highly representative of the different scientific communities.[5]

Another interesting (but sometimes neglected) aspect of institutional repositories is that they can collect, index, keep and disseminate grey literature products. The availability of certified data about those products

[1] Organisation for Economic Co-operation and Development (OECD), *Frascati manual 2002: proposed standard practice for surveys on research and experimental development: the measurement of scientific and technological activities*, Paris, OECD - Organisation for economic co-operation and development, 2002.

[2] Research Excellence Framework 2014 (REF 2014), http://www.hefce.ac.uk/research/ref/; IRRA - Institutional Repositories and Research Assessment, http://irra.eprints.org/about.html; Leslie Carr, John MacColl, *IRRA (Institutional Repositories and Research Assessment): RAE Software for Institutional Repositories*, IRRA, 2006, http://irra.eprints.org/white/; *Open Access to research outputs: final report to RCUK*, LISU and SQW consulting, 2008, http://www.rcuk.ac.uk/documents/news/oareport.pdf; *Open Access to research outputs: annexes: final report to RCUK*, LISU and SQW consulting, 2008, http://www.rcuk.ac.uk/documents/news/oaannex.pdf.

[3] NARCIS - National Academic Research and Collaborations Information System, http://www.narcis.nl/; Elly Dijk, *NARCIS: linking CRISs and OARs in the Netherlands: A matter of standards and identifiers,* in *EuroCris Workshop on CRIS, CERIF and Institutional Repositories, CNR, Rome, 10-11 May 2010*, http://depot.knaw.nl/6365/.

[4] Confederation of Open Access Repositories (COAR). Working Group 2: Repository Interoperability, *The Case for Interoperability for Open Access Repositories. Version 1.0*, COAR, 2011, http://www.coar-repositories.org/files/COAR_Interoperability_Briefing.pdf.

[5] Maurits van der Graaf; Marjan Vernooy-Gerritsen (editor), *The European Repository Landscape 2008: Inventory of Digital Repositories for Research Output*, Amsterdam, Amsterdam University press, 2009, DOI 10.5117/9789089641908.

could provide new perspectives to science and technology phenomena investigation.[6] Actually, grey literature products could be used as a significant evaluation set both for bibliometric analysis and for investigations aimed at understanding science and innovation dynamics, change driving ideas, knowledge basis used in particular scientific developments, connections and communication patterns in particular disciplinary contexts.

In general, we think that cooperative systems facilitate the traceability of the different research product life-cycle phases and of the related metadata (versioning, persistent identification, etc.). The cooperative approach should be further extended within the scientific community to quality certification by adopting open and transparent peer-review processes (open peer review, open peer commentary, etc.).

# Open Access repositories in R&D information system: strategic role of cooperation

Open Access repositories, whose number has been steadily rising in recent years, are an important component of the global e-Research infrastructure.[7] The real value of repositories lies in the possibility of interconnecting them to create a network that can provide unified access to research outputs and be (re-) used by OA service providers, researchers' communities, management information systems (CRIS)[8], statistical information systems, bibliographic databases, etc.[9] However, in order to achieve this goal, a *multilevel* interoperability is needed. The purpose of this paper is to provide a broad overview of multilevel interoperability between Open Access repositories and other R&D information systems, identify the major issues and challenges that need to be addressed, stimulate the engagement of the repository community and trigger a process that will lead to the establishment of a cooperative network of R&D information management systems.[10]

Today, Open Access repositories are increasingly being used to collect, archive, and disseminate all types of research outputs such as research articles, conference proceedings, dissertations, data sets, working papers and reports.

Currently, research product data and metadata managed by OA and commercial repositories and databases are not used for official statistics due to several problems, such as the influence of the different national policies and strategies on the scientific production; the lack of a coherent framework of commonly agreed strategies; the different methods, tools and criteria used to collect data within the different public and private organizations; the lack of common classification criteria for product types, semantics and fields of reference; the insufficient reliability of data provided by the main bibliographic data bases (data base structure issues, lack of bibliographic & authority control tools, etc.); and more.[11]

The research process is an international and distributed endeavor, involving a variety of stakeholders such as scientists as authors and grant recipients, policy makers, research institutions, universities, publishers, and research funding agencies – each with their own set of interests. An international collaboration is needed between these stakeholders (actors) in order to develop cooperative and dynamic methodologies and processes for data and metadata quality control.

Interoperability is a pre-condition for a cooperative and widespread infrastructure of R&D information systems and for the value-added services and tools that can be built on top of the repositories.[12] The quality

[6] *Ivi*, p. 19-21.

[7] ERA Expert Group 7 - EG 7: Rationales for ERA, *Developing World-class Research Infrastructures for the European Research Area (ERA)*, Luxembourg, Office for Official Publications of the European Communities, 2008, DOI 10.2777/96979; European Commission, *Work Programme 2012 - FP7 - Capacities: Part 1: Research infrastructures*, European Commission, 2011, European Commission, 2011, http://ec.europa.eu/research/infrastructures/pdf/wp2012_research_infrastructures.pdf#view=fit&pagemode=none [6]

[8] Keith G. Jeffery, Anne Asserson, *Institutional Repositories and Current Research Information Systems*, New Review of Information Networking, 14, n. 2 (2009), p. 71-83, doi:10.1080/13614570903359357 [7].

[9] Confederation of Open Access Repositories (COAR). Working Group 2: Repository Interoperability, *The Case for Interoperability op. cit*.

[10] Wendy White, *Institutional repositories: contributing to institutional knowledge management and the global research commons*, In 4th International Open Repositories Conference, *Atlanta, Georgia, 18th – 21st May, 2009* [8], http://www.mendeley.com/research/institutional-repositories-contributing-to-institutional-knowledge-management-and-the-global-research-commons/; M. Vernooy-Gerritsen, G. Pronk, M. van der Graaf, *Three Perspectives on the Evolving Infrastructure of Institutional Research Repositories in Europe*, Ariadne, n. 59 (April 2009), http://www.ariadne.ac.uk/issue59/vernooy-gerritsen-et-al/.

[11] Organisation for Economic Co-operation and Development (OECD), *Frascati manual 2002*, *op. cit.*; Yoshiko Okubo, *Bibliometric Indicators and Analysis of Research Systems: Methods and Examples*, in *OECD Science, Technology and Industry Working Papers*, Paris, OECD Publishing, 1997, doi: 10.1787/208277770603; Maurits van der Graaf; Marjan Vernooy-Gerritsen (editor), *The European Repository Landscape 2008: Inventory of Digital Repositories for Research Output*, *Op. cit.*, p. 100-110.

[12] ERA Expert Group 7 - EG 7: Rationales for ERA, *Developing World-class Research Infrastructures for the European Research Area (ERA)*, *OP. cit.*; European Commission, *Work Programme 2012 - FP7 - Capacities: Part 1: Research infrastructures*, *Op. cit.*

of these services depends on the data provided by repositories/CRIS/other information systems and on the standardization of "quality control processes" (quality of data and metadata collection and management processes).

Given the quantity and complexity of the problems affecting what in a broad sense could be called the R&D international information system, it seems evident to us that the interoperability should be implemented not only at the technical level but also at the political and organizational ones by all the institutions involved in the creation, management and use of the information resources.

Data and metadata model standardization is necessary in order to enable efficient data exchange and to allow researchers to find the desired information in the different research management systems.

From a strategic view point, the development of common logical and organizational data and metadata models in the Scientific and Research System is important for:

- giving a simplified view to describe the specific area of interest;
- allowing for a better communication and multilevel interoperability between different information systems (Current Research Information Systems[13], Institutional Repositories, OA Service Providers, public and commercial Bibliographic databases, statistical databases, etc.);
- supporting information workflow management;
- supporting management and evaluation activities.

The aim of such cooperation should be the development of a common multilevel interoperability network and the first step should be a survey about policies and guidelines for organization and workflows, available data and metadata standards, cooperative bibliographic, authority control and subject access systems, formats and access conditions, data use and re-use patterns, in order to gain sufficient insight into the scale of interoperability problems. Only on such basis, that is actual options, effective solutions can be developed and deployed.

The multilevel cooperation is necessary at the following levels[14]:

- Political: effective initiatives are needed at the national and international levels to favor open access to research results achieved through public funding; those initiatives should address and harmonize the different R&D stakeholders' interests;
- Institutional: academic and research institutions should define institutional and operational policies and carry out effective and widespread advocacy actions in their reference communities.
- *"For institutional record-keeping, research asset management, and performance-evaluation purposes, and in order to maximize the visibility, accessibility, usage and impact of our institution's research output"* [15];
- Economic and legal: Open Access is not zero-cost. Economic strategies are needed to sustain open access to public research products, based on the "author/institution pay" model; on the legal side, the adoption of Creative Commons (CC) licenses should protect intellectual property rights while granting open access;
- Technical-organizational: standards and commonly-agreed guidelines (based on a cooperative approach) are needed to certify data and metadata quality;
- Technological: OA greatly benefits from the development and widespread adoption of open standards and protocols and from the development of modular, interoperable and open source-based platforms for the management and diffusion of digital contents.

[13] A Current Research Information System (CRIS) records the R&D (Research and Development) activity either funded by or carried out by an organization, or within a thematic or subject area. Typically it covers projects, people (expertise), organizational structure, R&D outputs (products, patents, publications), R&D events and R&D facilities and equipment.

[14] Alma Swan, *Sharing knowledge: open access and preservation in Europe: Conclusions of a strategic workshop - Brussels, 25-26 November 2010 - Report*, Luxembourg, Publications Office of the European Union, 2011, doi: 10.2777/63410.

[15] Institutional Self-Archiving Mandate – Definition - ROARMAP (Registry of Open Access Repository Material Archiving Policies), http://roar.eprints.org/.

# Green road: institutional and disciplinary archives

*"...Two roads diverged in a wood, and I --*
*I took the one less traveled by,*
*And that has made all the difference."*
(Robert Frost, The road not taken, 1920).

As a matter of fact, we could poetically say "*two roads to OA diverged in the wood of 'online scientific publishing*":

- the "golden road" of OA journal-publishing , where journals provide OA to their articles (either by charging the author-institution for refereeing/publishing outgoing articles instead of charging the user-institution for accessing incoming articles);
- the "green road" of OA self-archiving, where authors provide OA to their own published articles, by making their own eprints free for all.

In our opinion, the Green Road is the one that could bring more benefits to the scientific community.
One of the main research access/impact problem is that journal articles are not accessible to all potential users, causing a lack of potential research impact. The solution is making all articles really Open Access, granting a free, immediate and permanent online access to the full text of research articles for anyone, anywhere, webwide.
On the other hand we should consider the two roads to OA complementary, as well: the green road, representing the fastest and safest way to reach immediate 100% OA, might eventually lead to gold too.
In fact OA self-archiving is not self-publishing without quality control; nor it is meant to be scientific documentation for which the author could request payment and royalties (e.g. books or magazine/newspaper articles). OA self-archiving is bounded to peer-reviewed research, written only for research impact rather than royalty revenue[16].

The main consequence of a wider OA diffusion is that the whole society could benefit from a faster information spreading and from an accelerated research cycle through channels in which researchers can immediately satisfy their needs. It has been proved that OA articles have a significantly higher citation impact than non-OA articles. Only 5% of journals are gold, but over 90% are already green (the green light to self-archiving is possible and authorized to authors); yet only about 10-20% of articles have been self-archived. To reach easily the '100% OA' goal, self-archiving needs to be mandated by researchers' employers and funders, as U.K. and U.S.A have recently recommended, and universities play a significant role in that. It is crucial that both funders and universities/research-boards mandate Green OA self-archiving, as not all research is funded and repositories are successful in attaining a considerable percentage of self-archiving only where a mandatory policy has been issued and enforced.

The main benefit supplied by OA, in general, and Green Road, in particular, is that researchers can increase visibility, usage and impact of their own findings, as well as their chance to find, access and use results from other researchers. On the other hand, Universities co-benefit from the increased impact of their researchers, because it also gives an excellent return on the investment to research funders, such as governments, charitable foundations etc. Finally, publishers likewise benefit from the wider dissemination, visibility and higher journal citation impact factor of their articles, and Open Access can generate new metrics to be used for assessing and improving research impact.

[16] S. Harnad, Open Access research, JeDEM 3 (1): 33-41, 2011

# OA and grey literature valorization

Grey literature plays a significant role in the context of scientific documentation managed and diffused through Open Access archives, indexed and aggregated by the main service providers. Since the Seventh International Conference on Grey Literature at Nancy in 2006, GreyNet community started increasing its research activities relating to the OA effect on grey literature.

The adoption of open standards and OAI protocols by the International OpenGrey network facilitates the interoperability between OA repositories and OpenGrey (System for Information on Grey Literature in Europe). That's a first important step in developing cooperative networks for data and metadata certification.

The diffusion of the International Open Access initiative might certainly facilitate the development and coordination of cooperative networks, implementing sustainable processes and guidelines for:

- a better quality certification of grey literature products (open peer review, open peer commentary, etc.) and related metadata (adoption of common metadata standards and mappings, cooperative bibliographic and authority control, versioning, persistent identification systems, etc.);
- a better intellectual property protection especially for multimedia materials, containing a significant percent on Education, Learning and Professional Training (Creative Commons License is still weak). Moreover, a significant number of 'grey' production - as pre-prints, fact sheets, standards and working papers, committee reports, dissertation and Phd thesis - , still gets a discontinuous or null visibility due to intellectual property rights[17];
- a better information to users about copyright constraints (when and in which terms could I use it?);
- a wider access to research products, which can improve their visibility and impact.

Integrating Grey and Peer-reviewed literature often hosted in IR would enable a global view of the total available sources in a given scientific field, as well as an enhancement of research output measurements and metrics. Finally, it would also give increased researcher and affiliation visibility and (most importantly) better research outcomes.

# Quality control: strategy, methods, processes and tools

Bibliographic standards and authority control tools are not sufficient to assure data and metadata accuracy, completeness and consistency.
Quality management systems are needed to define processes for the production and management of data and metadata (Trusted Digital Repositories) [18], which imply commonly agreed organizational models[19].
Only a shared effort can guarantee:

- Quality certification of the main data and metadata production and management processes;
- Commonly agreed bibliographic and authority control tools for metadata certification[20];
- Highly customizable software solutions, based on open standards and platforms.

In our opinion, after defining policies, strategies, services[21], methodologies and processes, the cooperative effort should be focused on the design and implementation of technical and organizational solutions able to

[17] Most of the Italian Phd Thesis indexed in Opengrey are not published, yet. Moreover, BNI (National Italian Bibliography) currently reports and describes all Italian Phd Thesis, also not published: in fact this document type is subjected to legal deposit at the National Library of Florence (in accordance with DPR 30.10.1997, n. 387, art. 4)

[18] International organization for standardization (ISO), *Space data and information transfer systems. Open archival information system: Reference model. Standard ISO 14721:2003,*Geneva, ISO, 2003.

[19]David Giarretta, Henk Harmsen, Christian Keitel, *Memorandum of Understanding to create* a *European Framework for Audit and Certification of Digital Repositories*, http://trusteddigitalrepository.eu/Site/Memorandum%20of%20Understanding.html.

[20] Mauro Guerrini, *Gli archivi istituzionali: Open access, valutazione della ricerca e diritto d'autore*, Milano, Editrice Bibliografica, 2010, p. 33-60; Jung-Ran Park, *Metadata Quality in Digital Repositories: A Survey of the Current State of the Art*, Cataloging & Classification Quarterly, 47, n. 3-4 (April 2009), p. 213 – 228; Marieke Guy, Andy Powell, Michael Day, *Improving the Quality of Metadata in Eprint Archives*, Ariadne, n. 38 (2004), http://www.ariadne.ac.uk/issue38/guy/.

[21] DINI Working Group Electronic Publishing, *DINI Certificate Document and Publication Services - 2010: version 3.0*, march 2011,

support interoperability between the different R&D information Systems[22]. To achieve this goal it is important to:

- adopt a web service-based architecture (as in the JISC Information Environment Architecture);
- use open source software for information & content management systems (CRIS) and digital repositories (DSpace, E-prints, Fedora, JDIAM, Alfresco, etc.);
- use standard protocols and solutions for harvesting, aggregation, deposit, retrieval, cross-linking and context-sensitive linking (e.g. OAI-PMH and OAI-ORE[23], SRW - Search & Retrieve Web Service, SRU – Search & Retrieve URL Service, SWORD - *Simple Web-service Offering Repository Deposit,* Open URL[24], *etc.*);
- define an optimal set of context metadata, make sure these metadata are stored in CRISs and create automatic procedures for transferring these metadata to the repositories (CRIS-driven repositories – see also CERIF Metadata Model[25]);
- define common intermediary XML schemas for complex applications, in interoperable semantic and syntax context, for metadata interoperability, which allows flexible granularity[26];
- use interoperable record formats and syntaxes (e.g. SGML, XML, XML-RDF, XML-MARC, XML-MODS, XML-METS, etc.);
- use common standard models for web based interchange (e.g. RDF[27])
- participate to and leverage experiences from the cooperative development and use of Knowledge Organization Systems in the context of the semantic web (thesauri, classification schemes, subject heading lists and taxonomies, etc.)[28];
- enable citation metadata automatic detection within publications; work out/implement various multilingual controlled vocabularies (content international classifications) for the information objects in the Scholarly and R&D Information Domain (work out - or fill - the CERIF semantic layer)[29];

http://nbn-resolving.de/urn:nbn:de:kobv:11-100182800.
A certificate that describes the technical, organizational, and legal aspects (including interoperability) that should be considered in setting up a scholarly repository service.

[22] Magchiel Bijsterbosch, Foudil Brétel, Natasa Bulatovic Dale Peters, Maurice Vanderfeesten, Julia Wallace, *PEER. D3.1 Guidelines for publishers and repository managers on deposit, assisted deposit and self-archiving*, 2009, http://www.peerproject.eu/fileadmin/media/reports/D3_1_Guidelines_v8.3_20090528.Final.pdf.

[23] OAI-PMH protocol limits interoperability to the unqualified Dublin Core schema, thus "flattening" research evaluation or increasing noise with an oversimplified metadata management process. Keith G. Jeffery, Anne Asserson, *Institutional Repositories and Current Research Information Systems*, *Op. cit.*; Open Archives Initiative – Object Reuse and Exchange (OAI-ORE) – Defines standards for aggregation of compound digital objects, http://www.openarchives.org/ore/.

[24] Knowledge Exchange, *Guidelines for the aggregation and exchange of usage data*, http://wiki.surffoundation.nl/display/standards/KE+Usage+Statistics+Guidelines#KEUsageStatisticsGuidelines-GuidelinesfortheaggregationandexchangeofUsageData

[25] Keith G. Jeffery, Andrei Lopatenko , Anne Asserson, *Comparative Study of Metadata for Scientific Information: the place of CERIF in CRISs and Scientific Repositories*, 2002, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.5689.

[26] In many metadata environments, particularly in that of the digital library, the problems of complex and highly flexible generic schemas are as acute as they are in that of CERIF - Common European Research Information Format. A tension arises particularly between flexibility and interoperability: the more potential approaches to encoding are offered by a standard, the more problematic is the transfer of metadata to different information systems and its interpretation and processing by them. Despite its great power as an encoding mechanism for the complex metadata needs of research environments, the CERIF model remains relatively underused in the area of research information management. Its flexibility and fragmented architecture in particular can produce significant problems for implementers and reduce its interoperability unless such key components as its semantic infrastructure are standardized between institutions. These problems were experienced by developer communities of such standards and were solved by some by using the architectural mapping features of SGML/XML. Without this facility in XML, the solution advocated here can replicate its best features but also add more powerful, non-syntactic features, such as semantic control.
The strategy has been tested thoroughly in several live research information management environments and found to be generally workable: the only problems experienced have proved to be those inherent in the metadata scheme on which the mapping to CERIF was based. The results have proved it to form a good compromise which allows the use of a key standard (with the consequent benefits of wider interoperability) in conjunction with a constrained, project-specific and more easily implemented element set. The successful application of this methodology suggests that it may be beneficial in the wider area of digital library metadata in general, where several key metadata schemas are more easily implemented when constrained it this way.
Richard Gartner, *Intermediary schemas for complex XML applications: an example from research information management*, Journal of Digital Information, 12, n. 3 (2011), http://journals.tdl.org/jodi/article/view/2069/2086.

[27] Resource Description Framework (RDF) – A standard model for web-based data interchange, http://www.w3.org/RDF/.

[28] SKOS - Simple Knowledge Organization System is an area of work developing specifications and standards to support the use of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web, http://www.w3.org/2004/02/skos/.

[29] The CERIF Semantics is one component of the CERIF 2008 – 1.2 Full Data Model (FDM). It aims at recommending a standardized formal semantics to be applied in the wider context of Current Research Information Systems (CRISs) with CERIF as the underlying data model to supply the relevant entities and their relationships. The semantic component in this version presents the current core semantics; that is, the types and roles considered relevant in a research context between the involved core entities. Compared to its preceding version, this release provides a major upgrade with respect to the quantity of relevant terms. EuroCRIS – The European Organization for International Research Information, *CERIF 2008 – 1.2 Semantics*, EuroCRIS, 2010.

- define and use common research product categories and types (for example, CERIF – result–publication classification); [30]
- develop a cooperative bibliographic and authority control[31] system for Institutional Repositories and CRISs;
- develop cooperative multi-version control systems[32];
- extensively use specific unique and persistent identification codes:
    - for the different research product types (Handle, URN, DOI, Open DOI for dataset[33], SICI, ISBN, ISRN, ISTC, etc.);
    - for the researchers (international author ID, ORCID[34], etc.);
    - for research information space, CERIF entities being the core;
    - for institutions and projects (international Digital Institution Id – DII - and international Digital Project Id - DPI);
- develop a cooperative Persistent Identifiers (PI) resolution system (meta-resolver for PI)[35];
- develop cooperative semantic and meta search and discovery systems and tools[36].

# CNR IA: a viable solution

In this section we will describe the situation of CNR research product archives, the current initiative aimed at implementing an Institutional Archive of research products and viable solutions to accomplish this task. A brief description of the CNR library system is given below, in order to allow a better understanding of the IA discussion.

CNR's library infrastructure reflects CNR's organization, featuring a Central Administration in Rome and a Scientific network made up of thematic institutes distributed all over the national territory. A significant percentage of CNR's institutes are hosted inside territorial Research Areas, which provide common services thus increasing efficiency.

CNR's library system features a hierarchical and distributed organization, which includes a Central Library (Biblioteca Centrale), Research Area Libraries (Biblioteche delle Aree di Ricerca), Institute Libraries (about 80). It provides a wide range of services to the entire scientific community and has recently adopted new organizational measures in order to increase the coordination of its different branches and improve the quality of the services provided to the internal scientific community. This effort has already produced some results in terms of process rationalization and digital resource sharing. The medium term objective is to complete the integration between CNR's libraries and to provide new added value services both to the internal and external scientific community.

At present, within our institution there are some research product archives but an Institutional Archive is not available. The existing repositories are based on open source platforms and are all OAI-PMH enabled.

An ad hoc working group has been established in order to define the architecture, standards, workflows and rules of a unified Institutional Archive. This group includes the personnel which has been involved in the development and management of the existing archives. The new architecture will be based on open standards and open source platforms. Web service interfaces will be provided for the communication with other systems.

From the researchers' perspective, auto-archiving will be implemented and favored. Obviously several levels of control will be enforced, in order to assure content and metadata quality. To this end, we think that the whole CNR library system should be involved, in order to have a first formal control at the local level (institutes and research areas) and a second one at the central level (Central Library). On the other hand,

[30] EuroCRIS – The European Organization for International Research Information, *CERIF 2008 – 1.2 Semantics*, *Op.cit.*

[31] VIAF – Virtual International Authority File, http://viaf.org/.

[32] Version Identification Framework Project, http://www2.lse.ac.uk/library/vif/index.html; VERSIONS (Versions of eprints. A user requirements study and investigation of the need for standards), http://www2.lse.ac.uk/library/versions/; The RIVER Scoping Study on Repository Version Identification - Sally Rumsey, Frances Shipsey, Michael Fraser, Howard Noble, Mark Bide, Hugh Look, Deborah Kahn, *Scoping Study on Repository Version Identification (RIVER) - Final Report*, 2006, http://www.jisc.ac.uk/uploaded_documents/RIVER%20Final%20Report.pdf.

[33] DataCite, http://www.datacite.org/.

[34] ORCID – Open Researcher and Contributor ID, http://orcid.org/.

[35] PersID – Project aimed at building a persistent identifier metaresolver infrastructure for digital publications and electronic resources, http://www.persid.org/.

[36] Kathleen Menzies, Duncan Birrell and Gordon Dunsire, *New Evidence on the Interoperability of Information Systems within UK Universities*, Lecture Notes in Computer Science, 6273 (2010), p. 104-115, DOI: 10.1007/978-3-642-15464-5_12.

quality control will be automated where possible, leveraging the quality control strategies, methods, processes and tools described in the previous sections.

One of the main benefits for researchers will be the possibility to produce certified lists of their own publications (e.g. for internal career advancement procedures). We think that this could be a good incentive for self-archiving.

# IA integration with CNR IS

Thanks to the web service based interfaces, the new system will be integrated with CNR Information system. Figure 1 shows the high level architecture of CNR IS. The new Institutional Archive is positioned in the right bottom corner.

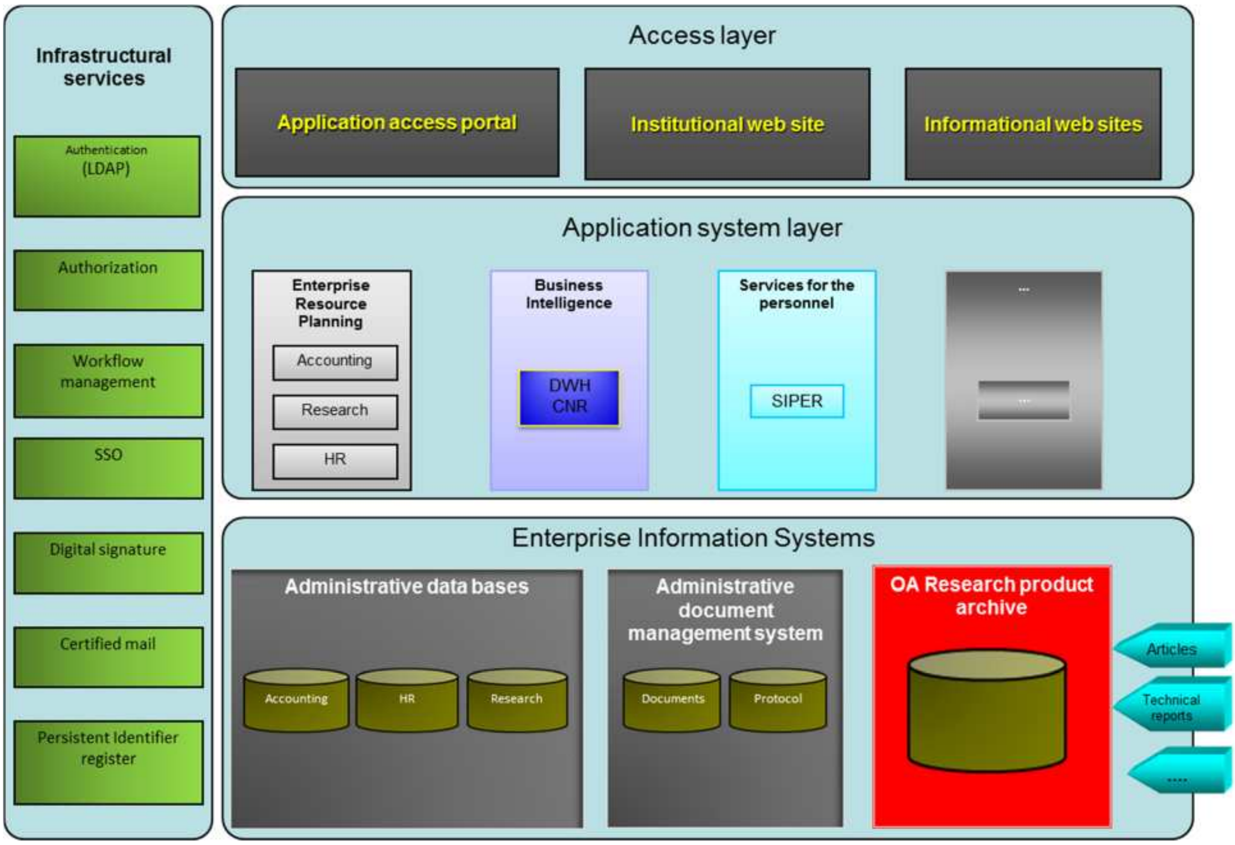

**Figure 1: CNR IS  high level architecture**

At the bottom of this architecture there is the Enterprise Information System layer, which includes the administrative data bases and document management systems. The new IA will be positioned at this level. The Application System layer includes all the systems and applications that manage or analyze the data kept at the underlying level. The Access layer includes all the portals and websites that provide access to services and information residing in the Application layer. Orthogonal to the described layers there is the Infrastructural Services one, which provides cross-application services to the entire IS, such as authentication, authorization, single sign on, etc..

Particular care will be put in implementing an actual interoperability of the new IA with other internal and external systems. The reference schema for interoperability will be the EuroCRIS one, described in Figure 2 (single institution) and Figure 3 (inter-institution interoperability).

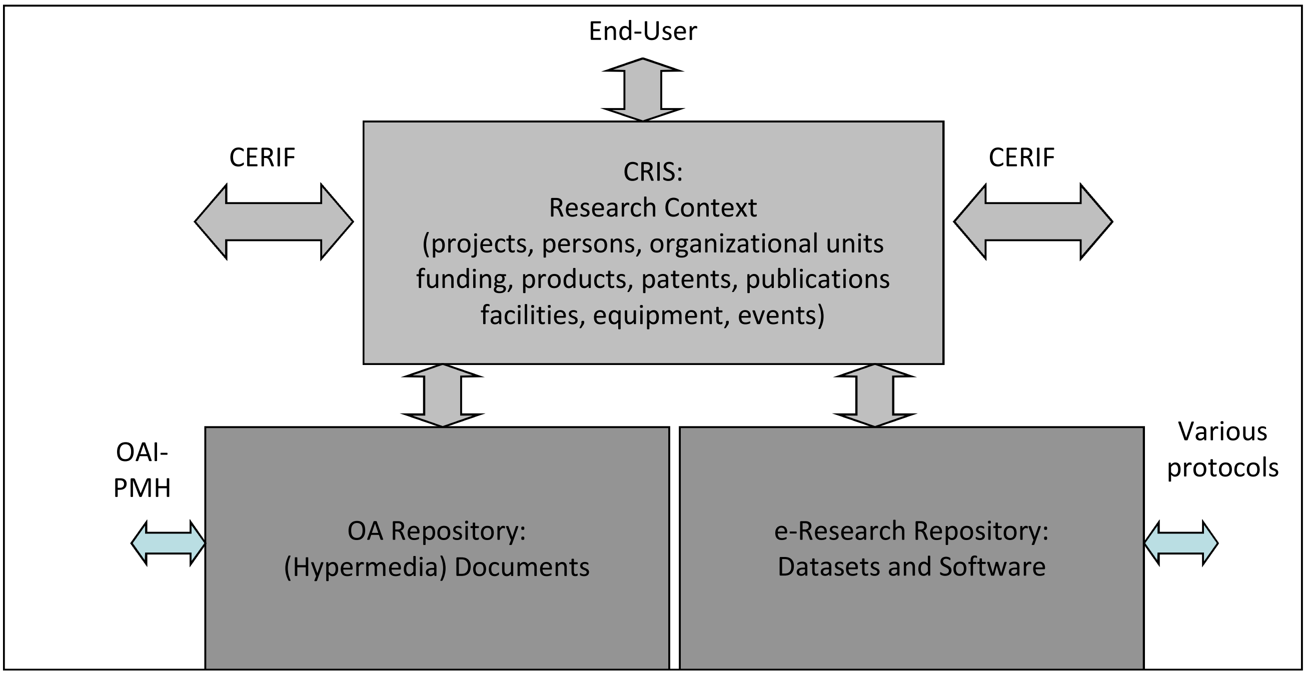

**Figure 2: EusroCRIS schema - Architecture for a single institution**[37]

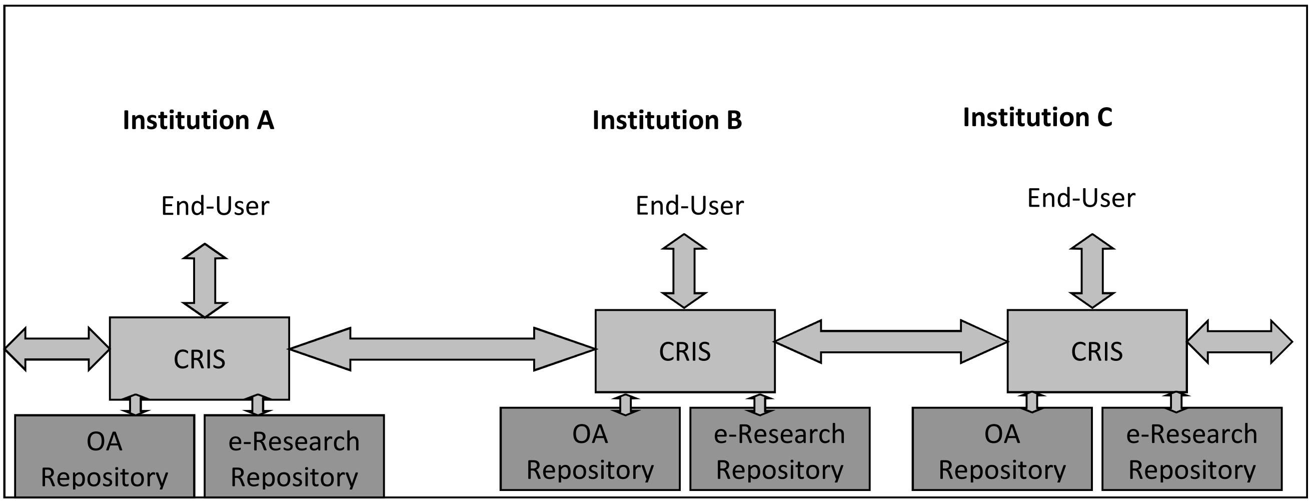

**Figure 3: EuroCris schema for CRIS and OA/e-research repositories interoperability**[38]

As regards the communication between systems, the figures clearly show that OAI-PMH will play a significant role at the repository level whereas CERIF will be the standard of choice for inter-CRIS communications.
Last but not least, persistent identification of digital resources, authors, institutions, projects, etc. will be taken in due account as well as standards for product classification.

[37] Keith G. Jeffery, Anne Asserson, *Institutional Repositories and Current Research Information Systems*, *Op. cit.*
[38] *Ibidem*.
An architecture for providing a complete research information environment at an institution is presented. The linking together, at an institution, of a "OA repository of articles (that is a repository of publications deposited institutionally for toll-free open access in parallel with a peer-reviewed publication), a CRIS (to provide contextual information), and an OA repository of research datasets and software provides that institution with an information resource suitable for all the end-users and roles. Furthermore, the formalized structure of the CRIS allows a reliable workflow to be engineered which, in turn, encourages deposit of research outputs by reducing the effort threshold by using intelligent prompts or suggestions based on the information already stored and any constraints on permissible values of attributes. However the requirements of the end-user extend beyond the individual research institution or funding organization. The institutional CERIF-CRIS system can be linked to others because they have a formal structure and, hence, can be interoperated reliably and in a scalable way. This, in turn, provides a network of access to institutional OA repositories or e-research repositories linked to each institutional CRIS via the CERIF-CRIS gateways, enhancing and controlling the access using the CERIF-CRIS information as formalized, structured, and contextual metadata which is more detailed than DC and suitable for intelligent (machine-understandable) interoperation.

# Conclusions and future work

We think that it is important to be aware that the organizational and technical problems regarding multilevel interoperability are currently being discussed and addressed (or have been discussed and addressed in the past) in several other contexts[39], which are partly overlapping with the (digital) library community[40]:

- World Wide Web Consortium (W3C) (communities and working groups for interoperability);
- EuroCRIS – the European Organization for International Research Information (community for Current Research Information System interoperability)[41];
- the OAI (Open Archive Initiative) community (open archives and service providers based on harvested metadata according to the OAI-PMH, OAI-ORE protocols);
- institutional repositories/OA disciplinary repository networks (OpenAire, COAR[42], etc.);
- the Grey Literature Network Service and the OpenGrey - multidisciplinary European database;
- scholarly networks for Open Access publishing initiatives (SPARC - *Scholarly Publishing and Academic Resources Coalition, DOAJ* - Directory of Open Access Journals, OAPEN - *Open Access Publishing in European Networks,* etc.);
- Knowledge Exchange[43].

We think that we should learn from these communities and start with them discussions and common developments. The reason is not only the high similarity of data, services and ambitions, but also the fact that scientific products and data will be shared in all of these international contexts, thus requiring basic metadata to be produced only once, close to the source, and be re-used and augmented in other service contexts.

In our opinion, initiatives should be launched at the international level in order to:

- analyze new service scenarios/use cases for records and services or adapt existing ones;
- establish permanent cooperation for on multilevel interoperability involving R&D information system communities[44];
- establish international agencies or cooperative networks[45] for the definition and maintenance of commonly agreed workflow systems, principles, rules and vocabularies.

Within the Italian R&D system we are currently addressing the interoperability issue between the various information systems, also following the stimulus provided by recent laws and rules in the field of research evaluation. Within this context, OA archives are acquiring a great relevance thanks to their role of research product management systems and institutional data sources. In order to assure content reliability, a common effort is required for the development of cooperative certification systems.

# References

1. Organisation for Economic Co-operation and Development (OECD), *Frascati manual 2002 : proposed standard practice for surveys on research and experimental development : the measurement of scientific and technological activities*, Paris, OECD - Organisation for economic co-operation and development, 2002, ISBN 92-64-19903-9;
2. Carr, Leslie; MacColl, John, *IRRA (Institutional Repositories and Research Assessment): RAE Software for Institutional Repositories*, IRRA, 2006, http://irra.eprints.org/white/;
   *Open Access to research outputs: final report to RCUK*, LISU and SQW consulting, 2008, http://www.rcuk.ac.uk/documents/news/oareport.pdf ;
   *Open Access to research outputs: annexes: final report to RCUK,* LISU and SQW consulting, 2008, http://www.rcuk.ac.uk/documents/news/oaannex.pdf;

[39] Digital Archiving Consultancy, *Towards a European e-Infrastructure for e-Science Digital Repositories: a report for European Commission*, 2008, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/e-scidr.pdf.

[40] Digital Library Federation (DLF), http://www.diglib.org/; DL.org Community – Digital Library Interoperability, Best Practices and Modelling Foundations, http://www.dlorg.eu/.

[41] EuroCRIS – The European Organization for International Research Information, http://www.eurocris.org/.

[42] COAR – Confederation of Open Access Repositories, http://coar-repositories.org.

[43] Knowledge Exchange is a co-operative effort that supports the use and development of Information and Communications Technologies (ICT) infrastructure for higher education and research.

[44] EUROHORCs, (European Heads of Research Councils), http://www.eurohorcs.org/E/Pages/home.aspx; EuroHORCs and the European Science Foundation, *Vision on a globally competitive European Research Area and road map for actions to help build it*, EUROHORCs, 2008;

[45] Caroline Sutton, *Sharing knowledge: EC-funded projects on scientific information in the digital age: Conclusions of a strategic workshop - Brussels, 14-15 February 2011 - Report*, Luxembourg, Publications Office of the European Union, 2011, doi:10.2777/63780.

3. Dijk, Elly, *NARCIS: linking CRISs and OARs in the Netherlands: A matter of standards and identifiers,* position paper presented at the *EuroCris Workshop on CRIS, CERIF and Institutional Repositories, CNR, Rome, 10-11 May 2010*, http://depot.knaw.nl/6365/;
4. Confederation of Open Access Repositories (COAR). Working Group 2. Repository Interoperability, *The Case for Interoperability for Open Access Repositories. Version 1.0*, COAR, 2011, http://www.coar-repositories.org/files/COAR_Interoperability_Briefing.pdf;
5. Van der Graaf, Maurits; Vernooy-Gerritsen, Marjan (editor), *The European Repository Landscape 2008: Inventory of Digital Repositories for Research Output*, Amsterdam, Amsterdam University press, 2009, DOI: 10.5117/9789089641908 - E-ISBN: 9789089641908;
6. ERA Expert Group 7 - EG 7: Rationales for ERA, *Developing World-class Research Infrastructures for the European Research Area (ERA)*, Luxembourg, Office for Official Publications of the European Communities, 2008, DOI 10.2777/96979, ISBN 978-92-79-08312-9;

   European Commission, *Work Programme 2012 - FP7 - Capacities: Part 1: Research infrastructures*, European Commission, 2011, European Commission, 2011, http://ec.europa.eu/research/infrastructures/pdf/wp2012_research_infrastructures.pdf#view=fit&pagemode=none
7. Jeffery, Keith; Asserson, Anne, *Institutional Repositories and Current Research Information Systems*, New Review of Information Networking, 14, n. 2 (2009), p. 71-83, doi:10.1080/13614570903359357 – OAI Item Identifier: oai:epubs.cclrc.ac.uk:work/ 51773;
8. White, Wendy, *Institutional repositories: contributing to institutional knowledge management and the global research commons*, In 4th International Open Repositories Conference, *Atlanta, Georgia* 18th *- 21st May, 2009*, http://www.mendeley.com/research/institutional-repositories-contributing-to-institutional-knowledge-management-and-the-global-research-commons/;

   Vernooy-Gerritsen, Marjan; Pronk, Gera. Van der Graaf, Maurits, *Three Perspectives on the Evolving Infrastructure of Institutional Research Repositories in Europe*, Ariadne, n. 59 (April 2009), http://www.ariadne.ac.uk/issue59/vernooy-gerritsen-et-al/;
9. Okubo, Yoshiko, *Bibliometric Indicators and Analysis of Research Systems: Methods and Examples*, in *OECD Science, Technology and Industry Working Papers*, 1997/1, Paris, OECD Publishing, 1997, http://dx.doi.org/10.1787/208277770603;
10. Swan, Alma, *Sharing knowledge: open access and preservation in Europe: conclusions of a strategic workshop - Brussels, 25-26 November 2010 - Report*, Luxembourg, Publications Office of the European Union, 2011, doi: 10.2777/63410 - ISBN 978-92-79-20449-4;
11. International organization for standardization (ISO), *Space data and information transfer systems. Open archival information system: Reference model. Standard ISO 14721:2003,*Geneva, ISO, 2003;
12. Giarretta, David; Harmsen, Henk; Keitel, Christian, *Memorandum of Understanding to create* a *European Framework for Audit and Certification of Digital Repositories*, http://trusteddigitalrepository.eu/Site/Memorandum%20of%20Understanding.html.
13. Mauro, Guerrini; Capaccioni, Andrea (a cura di), *Gli archivi istituzionali: Open access, valutazione della ricerca e diritto d'autore*, Milano, Editrice Bibliografica, 2010, p. 33-60, ISBN 9788870756920, http://hdl.handle.net/10760/15609;

   Park, Jung-Ran, *Metadata Quality in digital repositories: a survey of the current state of the art*, Cataloging & Classification Quarterly, 47, n. 3-4 (April 2009), p. 213 – 228, DOI: 10.1080/01639370902737240;

   Guy, Marieke; Powell, Andy; Day, Michael, *Improving the Quality of Metadata in Eprint Archives*, Ariadne, n. 38 (2004), http://www.ariadne.ac.uk/issue38/guy/;
14. DINI - Deutsche Initiative für Netzwerkinformation. Working Group Electronic Publishing, *DINI Certificate Document and Publication Services - 2010: version 3.0*, march 2011, http://nbn-resolving.de/urn:nbn:de:kobv:11-100182800;
15. Bijsterbosch, Magchiel; Brétel, Foudil; Natasa, Bulatovic Dale Peters; Vanderfeesten, Maurice, Wallace, Julia, *PEER. D3.1 Guidelines for publishers and repository managers on deposit, assisted deposit and self-archiving*, 2009; http://www.peerproject.eu/fileadmin/media/reports/D3_1_Guidelines_v8.3_20090528.Final.pdf
16. Knowledge Exchange, *Guidelines for the aggregation and exchange of usage data*, http://wiki.surffoundation.nl/display/standards/KE+Usage+Statistics+Guidelines#KEUsageStatisticsGuidelines-GuidelinesfortheaggregationandexchangeofUsageData;
17. Jeffery, Keith; Lopatenko, Andrei; Asserson, Anne, *Comparative Study of Metadata for Scientific Information: the place of CERIF in CRISs and Scientific Repositories*, 2002, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.18.5689;
18. Gartner, Richard, *Intermediary schemas for complex XML applications: an example from research information management*, Journal of Digital Information, 12, n. 3 (2011), http://journals.tdl.org/jodi/article/view/2069/2086;

19. EuroCRIS – The European Organization for International Research Information, *CERIF 2008 – 1.2 Semantics*, EuroCRIS, November 2010; http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/Release_1.2/CERIF2008_1.2_Semantics.pdf;
20. Rumsey, Sally; Shipsey, Frances; Fraser, Michael; Noble, Howard; Bide, Mark; Look, Hugh; Kahn, Deborah, *Scoping Study on Repository Version Identification (RIVER) - Final Report*, 2006, http://www.jisc.ac.uk/uploaded_documents/RIVER%20Final%20Report.pdf;
21. Menzies, Kathleen; Birrell, Duncan; Dunsire, Gordon, *New Evidence on the Interoperability of Information Systems within UK Universities*, in Lecture Notes in Computer Science, 6273 (2010), p. 104-115, DOI: 10.1007/978-3-642-15464-5_12;
22. Digital Archiving Consultancy, *Towards a European e-Infrastructure for e-Science Digital Repositories: a report for European Commission*, e-SciDR, 2008, http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/e-scidr.pdf;
23. EuroHORCs and the European Science Foundation, *Vision on a globally competitive European Research Area and road map for actions to help build it*, EUROHORCs, 2008;
24. Sutton, Caroline*, Sharing knowledge: EC-funded projects on scientific information in the digital age: Conclusions of a strategic workshop - Brussels, 14-15 February 2011 - Report*, Luxembourg, Publications Office of the European Union, 2011, doi:10.2777/63780 - ISBN 978-92-79-20451-7.