Research Data in Current Research Information Systems

Joachim Schöpfel, Hélène Prost, Violaine Rebouillat

▶ **To cite this version:**

Joachim Schöpfel, Hélène Prost, Violaine Rebouillat. Research Data in Current Research Information Systems. 13th International Conference on Current Research Information Systems, CRIS2016, 9-11 June 2016, Scotland, UK, euroCRIS, Jun 2016, St Andrews, United Kingdom. 10.1016/j.procs.2017.03.030 . sic_01331537

HAL Id: sic_01331537

https://archivesic.ccsd.cnrs.fr/sic_01331537v1

Submitted on 13 Jun 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Research Data in Current Research Information Systems

Joachim Schöpfel, GERiiCO laboratory, University of Lille 3 (France) (corresponding author)

Hélène Prost, CNRS, Nancy (France)

Violaine Rebouillat, Dicen-IDF laboratory, Paris (France)

Abstract

The paper provides an overview of recent research and publications on the integration of research data in Current Research Information Systems (CRIS) and addresses three related issues, i.e. the object of evaluation, identifier schemes and conservation. Our focus is on social sciences and humanities. As research data gradually become a crucial topic of scientific communication and evaluation, current research information systems must be able to consider and manage the great variety and granularity levels of data as sources and results of scientific research. More empirical and moreover conceptual work is needed to increase our understanding of the reality of research data and the way they can and should be used for the needs and objectives of research evaluation. The paper contributes to the debate on the evaluation of research data, especially in the environment of open science and open data, and will be helpful in implementing CRIS and research data policies.

Introduction

In principio erat data, at the beginning was the data, with software systems to manage the research data. Back in the 1970s those systems were antecedents of the current research information systems (CRIS) designed to store and manage data about research conducted at an institution or organization and to extract useful knowledge for research management (Jeffery 2004). But, as Keith G. Jeffery stated, "the end-user should be able to obtain not only information on projects, persons and organizations and their patents, products and publications (…) but also the actual publications online with references to the data upon which the work is based and any associated software, instrumentation, methods and techniques" (p.83).

So far, research performance has mainly been measured in terms of publications, patents and funding. Open Science changes the game by introducing research data in the assessment process. In the era of e-Science and Big Data, research data are to be considered, in the words of the Vice-President of the European Commission responsible for the Digital Agenda, Neelie Kroes[1], as "fuel" for economy and science. How do CRIS capitalize on this fuel? How should they, and why? The following paper presents an update and offers some responses and perspectives to the question/issue of the rapport between research data and CRIS. In particular, it addresses three topics:

- What does "research data" mean? In other words, what exactly is (or should be) the object of evaluation?
- How are research data identified? Or how should research data be identified?
- What is the link between evaluation and long-term preservation?

The discussion is followed by some recommendations for further development of CRIS. Its approach is "value agnostic" – it does not ask whether evaluating research data is good or bad. It rather

[1] http://europa.eu/rapid/press-release_SPEECH-14-229_en.htm

assesses the way in which the evaluation is performed, the conditions under which it is performed, and the way in which it could be improved.

Methodology

The study is part of ongoing research and development on research data in social sciences and humanities at the University of Lille (GERiiCO research laboratory and academic library) and at the National Conservatory of Arts and Crafts in Paris (DICEN-IDF research laboratory). Our paper is based on a triple methodology:

1. **Literature review:** we identified about 30 recent papers on the link between research data and current research information systems, with Google Scholar, the Scopus database and the euroCRIS DSpace CRIS digital repository[2]. The state of the art covers issues such as metadata, funders' requirements and granularity.
2. **Survey on data repositories:** we analysed the social sciences and humanities data repositories in the re3data directory[3], regarding their compliance with the requirements of research evaluation.
3. **Studies on data management:** we re-analysed our own former and ongoing surveys on research data management (Schöpfel & Prost 2016, Rebouillat 2015), in particular regarding the typology of data resources and results, in order to gain complementary empirical evidence for the discussion on the object of evaluation.

We define CRIS together with the recent EUNIS study rather pragmatically as "informational system(s), built in-house or purchased from a vendor, dedicated to collecting, analysing, reporting, providing access and disseminating research and development (R&D) information", in contrast to institutional repositories, i.e. "digital collection(s) of research outputs (mainly publications and datasets) aiming to collect, preserve and disseminate the intellectual production of a higher education or research institution" (Ribeiro et al. 2015, slide 8).

Findings

1. The topic "research data" in CRIS studies

All papers and meetings in current research information systems talk about data. However, only a small number focus on research data especially since 2010, probably due to their growing importance in the context of cyberinfrastructure and open science. Some papers confuse data on research with research data and produce misunderstanding and ambiguity between information about persons, units and projects, and research data defined as "factual records (numerical scores, textual records, images and sounds) used as primary sources for scientific research, (…) that are commonly accepted in the scientific community as necessary to validate research findings" (OECD 2006). Vanhaverbeke et al. (2014) for instance define research data as "research-related data" on persons, projects and organizational units, opposed to "datasets" stored in open access repositories along with metadata and publications.

Research data have been defined in many different ways but there is little consensus. Following the OMB Circular 110[4], research data can be considered as "the recorded factual material commonly accepted in the scientific community as necessary to validate research findings." Re3data.org distinguishes between fourteen different types of data (archived data, audio-visual data,

[2] http://dspacecris.eurocris.org/
[3] http://www.re3data.org/
[4] https://www.whitehouse.gov/omb/circulars_a110#36

configuration data, databases, images, network-based data, plain text, raw data, scientific and statistical data formats, software applications, source code, standard office documents, structured graphics, and structured text) but admits that there are other categories in the 1,500 indexed repositories. We'll cite two examples:

- Lyon & Pink from the University of Bath define research data as "the data, records, files or other evidence, irrespective of their content or form (e.g. in print, digital, physical or other forms), that comprise a research project's observations, findings or outcomes, including primary materials and analysed data" (2012, p.3). They add that research data can take a variety of forms and cite results of experiments or simulations, statistics and measurements, models and software, observations e.g. fieldwork, survey results, interview recordings and transcripts, and coding applied to these, images from cameras and scientific equipment, and textual source materials and annotations.
- The CODATA-ICSTI Task Group on Data Citation Standards and Practices (2013) applies a "broadly inclusive" description of digital research data that "refers as well to forms of data and databases that generally require the assistance of computational machinery and software in order to be useful, such as various types of laboratory data including spectrographic, genomic sequencing, and electron microscopy data; observational data, such as remote sensing, geospatial, and socio-economic data; and other forms of data either generated or compiled by humans or machines" (CIDCR11).

The usual format of research evaluation systems like CERIF[5] distinguishes between persons, units and projects; with regard to the research output, they usually take into account publications, patents and other products. Research data are part of the latter. In the following, we will use the term "research data" as part of the research output and as a "systematic, partial representation of the subject being investigated" (OECD 2006).

Compared to the content of the euroCRIS DSpace CRIS digital repository with 385 items published between 2002 and 2015, the sample of 27 papers dealing with research data in CRIS represents only 5% (figure 1). Two papers were published early in 2002 and 2004, while the main bulk was published more recently, over the last four years.

Today, the CRIS is evolving in a new environment of repositories and research data programs (Jeffery 2012). Nevertheless, while people and publications are crucial elements of CRIS, research data are not or less. A recent definition of key performance indicators (KPI) enumerates staff, PhD projects, research projects, funding and publication output such as paper in journals, proceedings, books and book chapters, citations and IF top journal publications but omits datasets (Vanhaverbeke et al. 2014). The final report of the EUNIS – euroCRIS joint survey on European CRIS and institutional repositories reveals that only half of the CRIS (51%) provide functionalities for research data management and that only one out of five institutional repositories (18%) contain datasets (Ribeiro et al. 2015, 2016).

[5] Common European Research Information Format, see http://eurocris.org/cerif/main-features-cerif

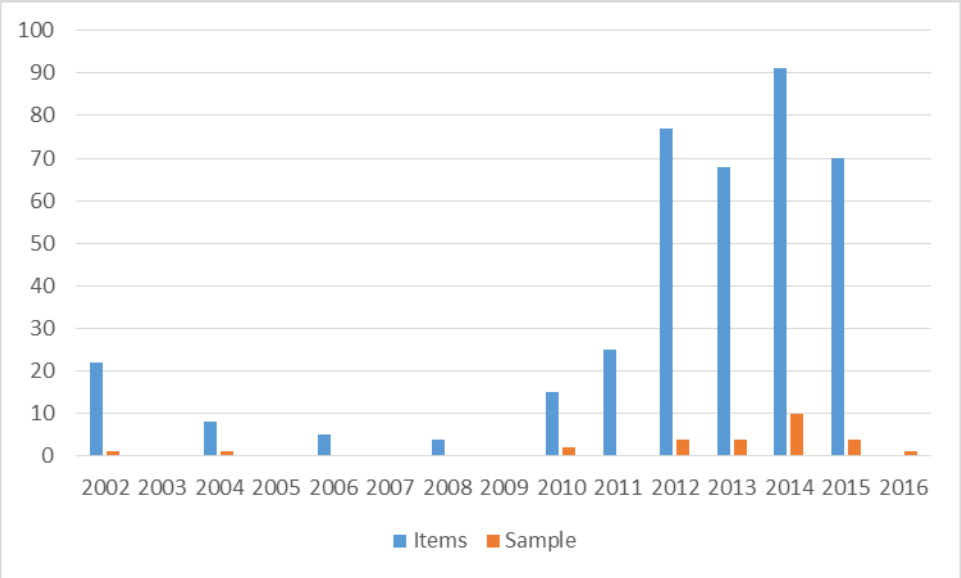

Figure 1: CRIS digital repository: items (n=385) and sample (n=27)

One reason may be that up to now the demand and usage of datasets stored in data repositories has remained limited. Hogenaar et al. (2010) state for the Dutch NARCIS platform: "(The) number of searches for datasets is remarkably small (7%)" (p.296). Another hindrance may be the lack of awareness, knowledge and/or skills by information professionals themselves. In their report on the implementation of Atira's PURE system[6] at King's College, London, McGrath and Cox (2014) observe that "another issue encountered was the need to validate more unusual forms of research outputs, e.g. web sites and datasets, which some library staff felt rather unqualified to perform".

In spite of this, CRIS are usually considered as an option to improve library services for research data management, especially by linking and storage. CRIS can enhance workflow and provenance control, due to identifiers, common vocabulary, rich semantics and links to publication, and improve data quality and efficiency (Doorn 2014). Clements and Proven (2015) highlight the potential of CRIS for the discovery of data underpinning research publications. Yet, while datasets are included in the general architecture of the University of St Andrews system (figure 2), they are not part of the CRIS process itself which lays emphasis on green and gold publications.

[6] Since 2012 Atira is part of Elsevier Research Intelligence.

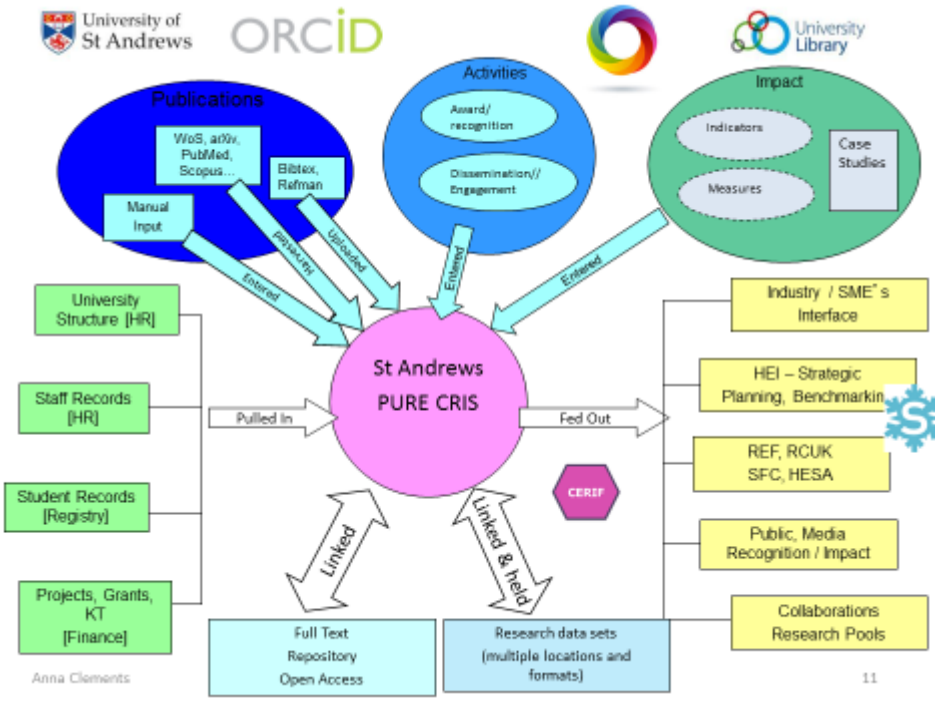

Figure 2: St Andrews CRIS architecture (Clements & Proven 2015, slide 11)

In other words, datasets are generally not considered as elements or objects of evaluation *stricto sensu*, in the same way as journal articles or patents. But CRIS can become a framework for the discovery of existing datasets, in order to foster access and make use of them.

2. Funding agencies' requirements

Funders hold the key for the development of research information systems. Doorn (2014) asserts that "research funding and performing bodies are taking an increasing interest in what happens to research data". They want credit to assert ownership, want to know about impact and reuse, and are interested in connecting data and publications (Ashley 2013). Data repositories are not only necessary to provide crucial evidence for publications, to allow data sharing and data reuse but they are moreover increasingly required by funding agencies for the deposit of datasets produced by funded research projects (Ribeiro 2013). These requirements constitute another reason for CRIS to integrate data.

Each funding agency labels its own and specific data-related requirements. Nevertheless, some essential criteria are common to most of them (Ashley 2013, Davidson et al. 2014, Doorn 2014):

- An explicit data policy (awareness) governing the research life cycle,
- A research data management plan,
- A statement about open access to the data,
- Long-term surely preservation (storage) for at least ten years,
- A structured metadata description.

The UK Engineering and Physical Sciences Research Council's (EPSRC)[7] expectations for instance include the requirements that organisations receiving EPSRC funding will:

- "Publish appropriately structured metadata describing the research data they hold (…).
- Ensure that EPSRC-funded data is securely preserved for a minimum of ten years (…).
- Ensure that effective data curation is provided throughout the full data lifecycle (…)" (Lewis 2014, p.2).

Often the attribution of a DOI is required, and sometimes the statement about open access is expected to include permanent access (accessibility). Other criteria are data monitoring, guidance for data management, linking publications to data, storage of non-digital research data (with conversion into digital format), the creation of a data repository and/or a data centre, data-related cost assessment, explicit access information (with conditions of restricted access), evaluation of impact and benefits (see for instance Lyon & Pink 2012).

The JISC funded DCC Discovery Service for UK Research Data promotes the RIF-CS interchange format designed for evaluation and compliance with CERIF (Ball et al. 2015). RIF-CS links datasets to projects (originating output), to persons (principal investigators), to publications (referencing datasets) and to other datasets (derivatives) and can moreover express the party that manages the dataset, the party that owns the dataset, a publication that cites the dataset, a publication that documents the dataset, and a publication of which the dataset is a supplement.

Increasing interest by funding bodies generally means strong recommendations or a code of conduct for the data management; it sometimes involves a clear and explicit request or mandate. Often, funding bodies expect the creation of a public data catalogue (metadata) and a data archive (preservation). Also, they expect institutions to provide the necessary human and technical infrastructure. Sometimes they offer additional funding for data curation. However, it is not always obvious how (and if) they will assure the follow-up, beyond the initial statement of intention.

3. Metadata (1): standards

In general, sufficient metadata and stable identifiers are considered as necessary to improve the usefulness of repository workflows (Littauer et al. 2012). Such a CRIS framework requires the incorporation of specific, data-related metadata which was for instance the goal of the JISC funded CERIF for the Datasets (C4D) project (Grinty et al. 2012). In a CRIS environment, a metadata catalogue constitutes a key component for the integration of data, data products and services (Bailo & Jeffery 2014), insofar as it provides description of datasets but also of software, services, users and resources like computers, data stores, laboratory equipment and instruments. Hodson (2013) notes a great diversity of metadata standards and stresses the importance of generic issues, such as place (geographic location, spatial coverage) and time (temporal coverage, time of production). Other topics are linked to population, licensing and access control. Accessibility plays a special role – metadata should inform if datasets are available for everyone or through selective disclosure.

For generic metadata, Ashley (2013) suggests international standards, such as the European INSPIRE directive on spatial information[8] for place or the General International Standard Archival Description ISAD(G)[9] for time. The Rutherford-Appleton Laboratory's Science Data Portal pilot project (Matthews et al. 2002) and the European OpenAIRE guidelines for data archives (Principe et al. 2014) are two

[7] https://www.epsrc.ac.uk/
[8] http://inspire.ec.europa.eu/
[9] http://www.ica.org/10207/standards/isadg-general-international-standard-archival-description-second-edition.html

major examples for a generic, domain agnostic metadata schema, reducing lack of consistency and providing interoperability through a limited number of properties. The OpenAIRE system links publications and authors to datasets, and then datasets are linked to data providers and projects (figure 3).

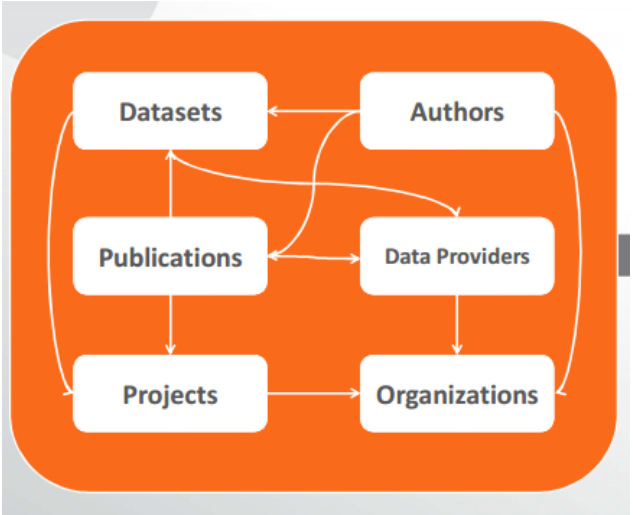

Figure 3: OpenAIRE integrated scientific information system (Principe et al. 2014)

According to the OpenAIRE guidelines, data repositories should contain datasets either as outcomes of funded research projects or linked with publications in the OpenAIRE information space. Their metadata should at least contain information on funding, access rights and licensing, related publications and datasets, and embargoes. The Dublin Core, as a kind of minimal standard for metadata of all kinds, appears compliant with research data, i.e. it can be used to describe these data; at least eleven out of the fifteen DC core elements can be qualified in a way that the DC supports functions required in a laboratory setting – discovery, usage, authentication, administration… (Bartolo et al. 2014). Matthews et al. (2002) suggest a complete mapping of all fifteen DC elements in the Science Data Portal metadata schema.

Citations, too, should support the discovery of data and their documentation, and they should facilitate the establishment of provenance of data. Citations and metadata are interdependent, as the CODATA report on ten emerging principles of data citation (2013) indicates. Citations "should employ widely accepted metadata standards" (p.6) as they "generally embed a limited number of metadata elements, such as a persistent identifier, descriptive title, and fixity information (for provenance verification). The data objects described by the citation are generally discoverable by this citation metadata (…)" (p.13).

4. Metadata (2): specificity and granularity

Beyond standards, enhanced metadata for datasets are typically specific to research fields, disciplines or institutions inferring a community perspective to data from different sources and disciplinary contexts, an approach that respects disciplinary workflows, tools and standards (Ashley 2013). Detailed information about datasets describe for instance particular subjects, data types, methods or scientific names (species). In other words, metadata should be as standard as possible,

but also as flexible as is needed, to accommodate the variant practices among communities, without compromising interoperability of data across communities (CODATA 2013).

One example for this mix of standardization and flexibility is the "RDE Metadata Profile for EPrints" developed during the Research Data @Essex project for the Essex Research Data Repository pilot at the University of Essex (Ensom and Wolton 2012), with 15 core and 31 "detail" elements, partly standard and controlled. This "mix" reflects the large heterogeneity of datasets, of their syntax (format), semantics (meaning) and schema (model). Koskela (2011) suggests a three-level architecture of descriptive, structural and administrative metadata (figure 4) designed to cope with this heterogeneity.

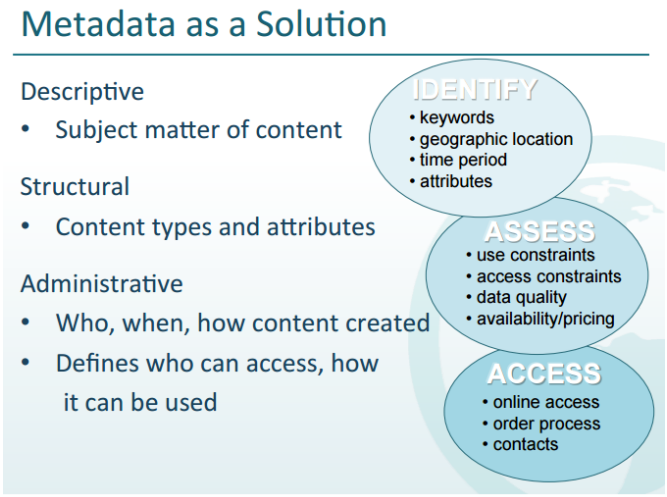

Figure 4: Three-level metadata architecture from DataONE (Koskela 2011, slide 9)

Granularity is a particular problem: some metadata catalogues provide a rather low level of detail and specificity in describing the various aspects of data and datasets (Elbaek et al. 2010). The Rutherford-Appleton pilot project proposes a model of scientific data holdings with two levels and three different items (figure 5).

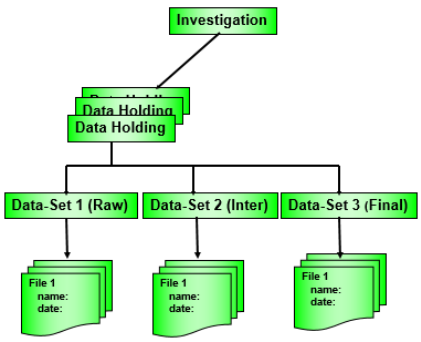

Figure 5: Hierarchy of data holdings (Matthews et al. 2002, slide 15)

According to this model, each data holding takes the form of a hierarchy – "one investigation generates a sequence of logical data sets, and each data set is instantiated via a set of digital files" (Matthews et al., p.197). These data files can be raw data, intermediary or final data. They are tied together by metadata but have different addresses and can be located on different servers. Their overall boundary depends on the initial investigation. Yet, data are not publications, their requirements are not the same as those of publications, they do not always have clean/clear boundaries, and their metadata must be able to support change – "changing data can have fixed metadata but don't force data to freeze" (Ashley 2013). Data curation throughout its lifecycle is *more* than preservation insofar as it implies dealing with change but also *less* insofar as it sometimes means data destruction (Ashley 2014).

The European Plate Observing System (EPOS) defines another data model with four data levels, compliant with the C4D project (Bailo & Jeffery 2014):

- "Level 0: raw data, or basic data (example: seismograms, accelerograms, time series, etc.)
- Level 1: data products coming from nearly automated procedures (earthquake locations, magnitudes, focal mechanism, shake-maps, etc.)
- Level 2: data products resulting from scientists' investigations (crustal models, strain maps, earthquake source models, etc.)
- Level 3: integrated data products coming from complex analyses or community shared products (hazards maps, catalogue of active faults, etc.)".

These four levels explicitly depend on the field of investigation but may be transposed to other domains, as a kind of domain agnostic data categorizations. Yet, the granularity remains a problem, especially of identifiers like the DOI which do not by themselves address granularity (CODATA 2013, CIDCR14). We will come back to this question below.

5. Data repositories and evaluation

In their case study on the University of St Andrews CRIS, Clements & McCutcheon (2014) highlight the interest of an authoritative list of trusted data repositories as a service for their researchers, for discovery and registry. Their problem has two names: heterogeneity of files, formats and metadata, and specificity of disciplines and instruments. A couple of years ago, a preliminary analysis of research data in the OpenDOAR directory revealed that only "a minority of archives used a specific variable to uniquely identify project ID, title and acronym (CERIF attributes of the core entity Project) and had specific functionality to make this information retrievable" (Luzi et al. 2012, p.81). At that time, OpenDOAR contained only 43 data archives (2% of all repositories), only twelve with project reference, i.e. with specific variables for projects or related information in other fields, providing the possibility to make the link. Luzi et al. concluded that an enhancement of data repositories was needed, using the CERIF model with its well-defined semantics. A recent paper defines four foundational principles to improve infrastructures supporting the reuse of research data, i.e. findability, accessibility, interoperability, and reusability (FAIR Data Principles, Wilkinson et al. 2016). These principles are useful to characterize well-curated data repositories through more or less stated criteria, like assignment of a globally unique and persistent identifier (F1), rich metadata (F2/R1), a standardized and open communications protocol (A1), a formal, accessible, shared, and broadly applicable language for knowledge representation (I1), or a clear and accessible data usage license (R1.1).

All three papers share the evidence that all data repositories are not qualified for research evaluation. In order to gain more empirical insight, we analysed the data repositories labelled by the international registry re3data.org with regard to selected criteria. We limited our sample to

repositories covering the social sciences and humanities, i.e. 413 repositories representing 27% of the re3data.org content. The survey was conducted in March 2016. The repositories are located in 43 different countries, mostly in the United States (38%), Germany (20%), the United Kingdom (14%) or in other Member States of the European Union. Nearly all are non-profit institutional data repositories (97%), with 66% of them having an explicitly disciplinary character.

Identifiers: Do the research data repositories use a persistent identifier system to make their provided data persistent, unique and citable? In our sample, 200 repositories don't (48%). 128 repositories use DOI (31%), 83 use handle (20%), while other identifiers (URN, ARK, Purl) are hardly used. Re3data.org offers very little information about author identifier systems; ORCID is used by only nine repositories (2%).

Metadata: Do the repositories apply metadata standards? Only 155 repositories do (38%). The most important standard is the Dublin Core (12%), followed by the Data Documentation Initiative (10%) and the W3C RDF Data Cube Vocabulary (2%). The part of domain agnostic, generic standards can be estimated at 16% (Dublin Core, W3C, DataCite, OAI-ORE) while the part of domain specific metadata schemas is 22% (social sciences, surveys, geography, ecology…). 156 repositories allow versioning (38%), in other words they take into account the dynamic character of datasets throughout the project and data lifecycle.

Quality: The directory does not inform about preservation policy. Yet, the existence of quality management can be considered as an indicator that a repository is committed to long term preservation, at least for five to ten years. 284 data repositories have some kind of quality management (69%). 136 repositories have obtained a certificate or label (33%), such as the Dutch Data Seal of Approval DSA, a domain agnostic label (14%), the German Council for Social and Economic Data RatSWD label (6%) or the European Common Language Resources and Technology Infrastructure CLARIN certificate B (4%).

Open access: 330 repositories (80%) provide complete open access to the deposited data. 65% can (also) provide restricted access (for registered users, institutional members, on demand, required fees), 10% can manage embargoes and 15% can support closed access.

Licensing: 156 repositories (38%) disseminate their data with full copyright but this does not mean that they do not allow some kind of licensing, on request. 137 repositories support the Creative Commons CC licenses (33%), 23 disseminate data under the CC Public Domain CC0 license (6%), 14 under the Open Data Commons Attribution License ODC (3%) and 11 under the Open Game License OGL (3%). Yet there are many other licenses, institutional or domain specific.

Only few repositories satisfy all conditions, i.e. are compliant with the main CRIS requirements. The re3data.org directory reveals a landscape of a large number of very different data repositories, domain or institution specific rather than generic. Variety and diversity prevail, which reflects the proximity with research communities and institutions but at the same time makes the assessment of their usefulness for evaluation and CRIS more difficult.

Our paper is limited to social sciences and humanities. However, even if some details may be different, this general observation applies also to the other fields of research, to scientific, technical and medical domains (STM). The part of STM data repositories in re3data.org without persistent identifiers is even higher (>70%) while the part with some kind of quality assurance is roughly the same (64%). There are fewer repositories with generic metadata standards such as the Dublin Core (only 2%), few repositories disseminate their data under CC licenses (<20%). But the overall statement remains valid: diversity and specificity is the rule, and the integration into a system

architecture built on standards, domain agnostic schemas and interoperability may be less easy than one could expect.

Three French examples from an ongoing survey on data repository management in France may illustrate the observations made above. The first example is ORTOLANG, a data repository in the field of linguistics[10]. ORTOLANG is a public, non-profit repository for text corpora, funded by the French Government (Pierrel 2014). The repository offers secured storage; one part of the 2000+ datasets is back upped for long-term preservation by CINES, the French Supercomputing and Digital Archiving Centre for Higher Education. ORTOLANG applies two persistent identifier systems, handle and ARK, admits versioning and is compliant with the Dublin Core metadata schema. Re3data.org indicates that ORTOLANG is part of the future French node of the European CLARIN network (see above) and cooperates with the DARIAH Digital Research Infrastructure for the Arts and Humanities. So far, the ORTOLANG is not certified CLARIN data centre. The Data Seal of Approval certification is in progress. The datasets (text samples) are standardised; they are disseminated in open access, partly under different Creative Commons licenses (CC-BY, CC-BY-NC-SA, International and French). Management of limited access, embargoes and confidentiality is possible.

The main challenge of ORTOLANG is the integration in the European network of data repositories in linguistics and digital humanities, which implies interoperability, quality assurance (maintaining a high service level) and standardization. The main problem will be its sustainability, i.e. the future funding of the repository and its staff beyond the project period (2013-2016/2019). For instance, limited funding is the main reason why only one part of the ORTOLANG datasets is transferred to CINES for long term preservation.

The second example is beQuali, a French data repository for qualitative surveys in social sciences hosted and maintained by the Paris University of Political Sciences, Sciences Po[11]. Like ORTOLANG, BeQuali is funded by the French Government (Duchesne & Garcia 2014). Launched in 2012, beQuali still has a project status, with only five datasets online (surveys). The repository offers long term preservation via CINES. BeQuali has no persistent identifier system; it adopts the metadata standard of the Data Documentation Initiative (DDI) compliant with the Dublin Core set of 15 elements[12]. BeQuali is part of the French network Quetelet, the French portal for data in the Humanities and Social Sciences[13], and a member of CESSDA, the Consortium of European Social Science Data Archives. So far, beQuali is not certified or labelled. The survey metadata are available in open access, including detailed information about the genesis and realization of each survey. The datasets themselves, i.e. the transcriptions of interviews, the guidelines and survey reports are available on duly substantiated request only, under the normal legal regime (copyright). Management of limited access, embargoes and confidentiality is possible.

The main challenge of beQuali is awareness and deposits. The main problem will be future funding, after the end of project funding. Interoperability, standardization, licensing and international networking have not appeared to be priority issues so far.

[10] https://www.ortolang.fr see also re3data.org: Ortolang; editing status 2015-04-02; re3data.org - Registry of Research Data Repositories. http://doi.org/10.17616/R3T054 last accessed: 2016-04-17
[11] http://bequali.fr/fr/
[12] http://www.ddialliance.org/
[13] http://www.reseau-quetelet.cnrs.fr/spip/?lang=en

The third and last example is from STM: SEANOE[14], a non-profit research data repository in Marine Science launched in 2015 and funded by the French Public Research Institute for Exploitation of the Sea Ifremer (Merceur 2015). In April 2016, SEANOE published more than 100 datasets, each with a DOI. The metadata schema is compliant with the Dublin Core. The long-term preservation of data filed in SEANOE is ensured by Ifremer infrastructure. All datasets are published in open access and under Creative Commons licenses. An embargo of up to 2 years is possible for example, to restrict access to data of a publication under scientific review. SEANOE is not certified so far. Together with other French public research organisations Ifremer is preparing a quality label "Pôle Océan" for the certification of marine science data[15]. On this level, Ifremer is cooperating with several international programmes and networks, including the Research Data Alliance[16].

The main challenge for SEANOE is NOT funding or long-term sustainability because for many years now Ifremer has been conducting a pro-open access policy and is funding open access initiatives and infrastructures, such as the institutional repository Archimer[17] and now the data repository SEANOE. In 2010, Ifremer decided on a total mandate for publications[18]. Regarding data, Ifremer has no mandate but an explicit goal, to make public research results available (open science), reproducible and citeable in a secured environment. So the main challenge of SEANOE is raising awareness and fostering uptake and acceptance by the French Marine Science community, through a mix of communication and user-oriented development, to increase straightforward handling and monitoring.

One size does not fit all, and our idea is NOT that all data repositories should follow the same schema. However, some internal and external (situational) factors seem helpful in achieving compliance with requirements of CRIS and research evaluation, such as international networking, an explicit pro-open access (open science) policy, leadership by information professionals and partnership with a digital preservation service provider. On the other hand, limited funding and lack of sustainability are threats to this compliance.

6. Research data management

Our recent studies on research data management (Schöpfel & Prost 2016) and research data in dissertations (Prost et al. 2015), together with our work with PhD students on data management and data sharing, add more empirical evidence on the topic. The main issues are:

Output/input: The distinction between primary and secondary data is essential for research data management and data sharing. However it is unclear what should be measured for research evaluation. Normally, only output should be assessed - in other words, secondary data. On the other hand, collecting and curating primary data as input for research is time and resource consuming and co-determines the quality of the scientific result. Should input be valued? Assessed? How? As a resource or as a result, i.e. output? Primary data may also be third-party data "which may have originated within the institution or come from elsewhere (sourced) for re-use (as an input) as part of research projects (…) with terms and conditions specified by the data owners (e.g. in contracts, licences, re-use agreements)" (Lyon & Pink 2012, p.3). Should those data initially produced by others be considered as output?

---

[14] http://www.seanoe.org/ see also re3data.org: SEANOE; editing status 2016-01-26; re3data.org - Registry of Research Data Repositories. http://doi.org/10.17616/R3J33X last accessed: 2016-04-19

[15] http://www.pole-ocean.fr/en/The-Pole-Ocean

[16] http://www.pole-ocean.fr/en/The-Pole-Ocean/International-context/International-programmes

[17] http://archimer.ifremer.fr/

[18] http://roarmap.eprints.org/146/

Generic/specific data: Our surveys show a great variety of research data, photographs, spreadsheets and databases, text and speech samples, surveys, experimental data, interviews and so on. Some of them are more or less specific to one or two disciplines while others (most) are more largely distributed among the different domains of social sciences and humanities. According to our survey results, the data distribution ("data profile") seems to characterize a discipline better than a specific type of data. This is a strong argument for a domain agnostic approach to metadata and evaluation.

Functional/dysfunctional: What researchers do with "their own data" is more or less functional and compliant with their immediate needs, albeit they generally acknowledge the lack of back-up and long term preservation solutions. In many cases, this form of data curation remains highly individual, local and sometimes even private, e.g. when research data are stored on the hard disk "at home". This functional practice becomes dysfunctional regarding CRIS requirements and evaluation, which imply collective, standard and community procedures, not private practice.

Funders are the key (but not exclusively): Funders' requirements are a strong incentive for research data management and sharing. H2020 guidelines on project proposals are a very (if not the most) important motor for preparing a data management plan. Just as important are requests from other researchers or the own institution (research centre, university).

Discussion – evaluation, identifiers and preservation

What exactly is data? What should be evaluated? How can one distinguish between collection of data as input (primary data) and produced data as output (secondary data)? At what level of granularity should the research data be evaluated? Which typology should be applied? Can we identify any attempts of an evaluation-relevant typology of data?

As we stated above, there is no widely accepted definition of the term or concept of research data. The Royal Society considers data as "qualitative or quantitative statements or numbers that are (or assumed to be) factual" (2012, p.104). Just as data produced during and through scientific activity, research data are a sub-category. Their characteristics and description depend largely on their discipline, on instruments or procedures of collection and recording, also on their processing in order to become exploitable ("readable") or for exploitation.

Thus, the Research Information Network suggests a data classification based on the mode of production or generation (observational, experimental, models or simulation, derived or compiled, reference or canonical)[19]. The re3data repository distinguishes between fourteen different types or formats of data but admits that the indexed repositories contain lots of other formats. A third distinction can be made following the content of datasets (sociological surveys, DNA nucleotide sequences, algorithms…). But because of the rapid development of new formats and contents, these models are all but stable or exhaustive. Also, if data is defined too narrow, too specifically related to a discipline or field of investigation, how can it be evaluated and compared to others? In their "Roadmap for EPSRC" for the University of Bath, Lyon & Pink (2012) insist on the fact that research data should be abstracted from the subjects of research, and as such should not include the subjects of research themselves.

Object of evaluation

In summary, in line with empirical studies and research on CRIS, evaluating research data as a specific part of scientific output together with scientific publications, gives attention rather to secondary data and apply a generic approach to typology and methodology (procedures). But this approach remains essentially descriptive and does not assess whether a given dataset produced by research project A is

[19] See the description on the University of Bristol website https://data.bris.ac.uk/bootcamp/data/

"better" (whatever this means[20]) than another dataset from research project B. Also, it does not say anything of the potential impact of the produced data.

The development of CRIS did not produce any special data typology. In fact, as our literature review shows, current research information systems generally do not evaluate research data in the strict sense but data management, along with a more or less generic and standard description. This is quite different from publications which are categorized, counted and pondered to provide institutional or individual metrics. Regarding research data, even if they are considered as scientific output, they are not assessed or measured but only described. CRIS usually assess investment in data management and evaluate formal or procedural criteria like the existence of a data management plan (stewardship), the application of data standards, citability, the sharing of datasets (open access), copyright, third-party right and ethical clearance, reusability, identification and long term preservation.

We asked above: "What exactly is data? What should be evaluated?" In the context of CRIS, the answer may seem paradoxical – the exact nature of data seems more or less irrelevant for research evaluation, because the real object of evaluation is data curation NOT data. Two other aspects appear of particular interest, i.e. identifiers and preservation.

Identifiers

Persistent identifiers are a crucial condition for the identification of datasets and their linking to publications, to make workflow URLs permanent and prolong workflow longevity long after publication (Littauer et al. 2012). Our survey on data repositories shows that the DOI is usually the preferred solution, compared to other options e.g. handle, URN etc. For instance, the French Research Institute for Exploitation of the Sea (IFREMER) decided recently to assign DOIs to all research data, in order to improve the visibility of their oceanographic projects and facilitate research evaluation[21]. The same decision was taken at St Andrews: "Extending the CRIS to include research data can be achieved by adding metadata to the CRIS with external links where appropriate e.g. DOI to the data itself" (Clements & McCutcheon 2014, p.4).

But at what level of granularity should the DOI be assigned? A high-level DOI makes the access to data more difficult. A low-level DOI makes the citation difficult. CODATA (2013) suggests that "citations should support the finest-grained description necessary to identify the data (...) (However) the optimum level and nature of granularity, however, would vary with the kind of data" (CIDCR16). What constitutes a whole data set does not always seem obvious (Duke & Ball, 2012). A dataset, so the CODATA report, may form part of a collection and be made up of several files, with each containing several tables and many data points. It is possible but not necessary to attribute different DOIs to the same dataset, on different levels. This manifest problem of the so-called granularity principle makes it difficult to use identifiers for the evaluation of research data. DOIs are useful for discovery, citation etc. but not for evaluation of data. In fact, CRIS evaluate the assignment of DOIs or other persistent identifiers as a criterion of high-quality research management NOT the datasets identified through DOI themselves.

Preservation

Our last question concerns the preservation issue – how is long-term conservation related to evaluation? Generally, studies on data curation and CRIS consider long-term conservation necessary and critical for research evaluation. The critical issue can be described as the choice of the

[20] More investment (budget, human resources)? A greater volume (file size, number…)?
[21] See http://www.ifremer.fr/sismer/index_UK.htm

repository(ies) for the deposit of the research data produced by an institution, a research project or an individual scientist. The objects of assessment vary from policy statement (declaration, guarantee) and institutional investment (human resources, budget) to certification, quality label and institutional partnership with another expert and specialised organisation (service) in charge of data preservation (data centre…). The re3data.org directory reveals that only one part of the data repositories meets the needs of research evaluation formats and systems.

Should long-term preservation include data sharing, open access to research data? Usually, open data policy is part of research evaluation, yet dependent on several conditions such as third-party rights, confidentiality, etc. The European Amsterdam Call for Action on Open Science imposes on the Member States the fact that open data should set the default standard for publicly funded research and that standardised data management plans should become an integral part of the research process and a precondition for funding. Open access should be the default, but other access regimes are allowed, from open and free downloads to application and registration-based access. "Conditions can be dependent on the nature of the data, common practice within a specific academic discipline, legal (privacy) frameworks, and legitimate interests of the parties involved"[22]. So again, CRIS usually appear assessing rather the (formal) possibility and/or policy of open research data rather than the real openness.

Conclusion

As research data become a crucial topic of scientific communication and evaluation, current research information systems must be able to handle them. Our paper provides an overview on recent research and publications on the integration of research data in CRIS, with a focus on social sciences and humanities. The literature review shows that only a small number of studies focus on research data. While people and publications are crucial elements of CRIS, research data appear less important. Nevertheless, CRIS are usually considered as an option to improve library services for research data management, especially by linking and storage. CRIS can enhance workflow and provenance control, improve data quality and efficiency and facilitate the discovery of data underpinning research publications.

The Amsterdam Call for Action expects National authorities and Research Performing Organisations to put in place an institutional data policy which clarifies institutional roles and responsibilities for research data management and data stewardship. Funders are taking an increasing interest in what is happening to research data, they want credit to assert ownership, they want to know about impact and reuse, and are interested in connecting data and publications. Often, funding bodies expect the creation of a public data catalogue (metadata) and a data archive (preservation). Also, they expect institutions to provide the necessary human and technical infrastructure.

While the need for generic and domain agnostic, standard metadata and stable identifiers are put forward, namely for place, time, linking, licensing and accessibility, enhanced metadata for datasets are typically specific to research fields, disciplines or institutions inferring a community perspective to data from different sources and disciplinary contexts, an approach that respects disciplinary workflows, tools and standards.

Further empirical evidence shows that only few data repositories are compliant with the main CRIS requirements, e.g. standard metadata, identifiers, long term preservation, etc. Variety and diversity prevail, which reflects the proximity with research communities and institutions but at the same time makes the assessment of their usefulness for evaluation and CRIS more difficult. Studies on research

[22] http://english.eu2016.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science

data management reveal that even if the distinction between primary and secondary data is essential for research data management and data sharing, it is unclear what should be measured for research evaluation. They provide evidence, too, for the interest of a domain agnostic approach to metadata and evaluation, for the potential conflict between individual information behaviour and CRIS requirements, and for the strong influence of funding bodies and also of institutions.

Especially when considering the funding bodies' requirements and their translation into CRIS functionalities, it appears that generally they do not evaluate research data in the strict sense but data management, along with a more or less generic and standard description. The object of evaluation are data related criteria, formal aspects e.g. data management plan, assignment of DOI to datasets, rich and standard metadata (indexing), deposit in a labelled data repository (DataSeal, FAIR…), liberal dissemination in the context of Open Data and Open Science.

Now, these data-related criteria are often in charge of information professionals, e.g. academic librarians or data officers, rather than of scientists. They are part of project management and governance, perhaps even of equipment (resources). It should be considered whether in the CRIS environment, research data is really "output" or "product" like articles or patents, especially because of the difficulty to define the data beyond mere description, to distinguish between primary and secondary data, and to determine its granularity. At least, data, i.e. data management should not be part of evaluation of researchers but only of projects or institutions.

A minimal list of recommendations for integration of data in research evaluation would cover at least six aspects:

- Evaluation should not concentrate on data but on data management.
- The deposit of data in labelled data repositories should be preferred (expected).
- Standard, generic and rich metadata should be required.
- Standard persistent identifiers for data and contributors (authors), namely DOI and ORCID, should be required.
- Open data policy should be the default, at least for public funded research.
- Evaluation should include explicit measures for reporting and follow-up (no simple declaration of intention).

It may be too early to provide definitive answers to all questions, and more studies on the evaluation of research data will be needed, in particular about granularity, metadata, licensing (accessibility) and preservation. However, as a conclusion to our improved understanding of research data we can already suggest that the future development of CRIS software and CERIF should be careful with the issue of research data and consider how this reality is compliant with the CERIF data model. At the same time, the requirements of both funding bodies and CRIS should contribute to further standardization and improvement, in terms of content, quality and certification of data repositories in order to enhance their usefulness for research evaluation.

References

Ashley, K., 2013. (Research) dataset metadata – requirements. In: *11th euroCRIS Strategic Seminar: "Metadata in Research Information Systems",* Brussels, Sep 9-10, 2013. http://dspacecris.eurocris.org/handle/11366/272

Ashley, K., 2014. My data, our data, your data: data reuse through data management. In: *CRIS2014: 12th International Conference on Current Research Information Systems*, Rome, May 13-15, 2014. http://dspacecris.eurocris.org/handle/11366/320

Bailo, D., Jeffery, K. G., 2014. EPOS: a novel use of CERIF for data intensive science. In: *CRIS2014: 12th International Conference on Current Research Information Systems,* Rome, May 13-15, 2014. http://dspacecris.eurocris.org/handle/11366/185

Ball, A., Brown, C., Molloy, L., Van den Eynden, V., Wilson, D., 2015. Using CRIS to power research data discovery. In: *euroCRIS Membership Meeting 2015 – Spring*, AMUE, Paris, May 11-12, 2015. http://dspacecris.eurocris.org/handle/11366/378

Bartolo, L. M., Lowe, C. S., Melton, A. C., Strah, M., Feng, L., Woolverton, C. J., 2014. Effectiveness of tagging laboratory data using Dublin Core in an electronic scientific notebook. In: *CRIS2014: 12th International Conference on Current Research Information Systems*, Rome, May 13-15, 2014. http://dspacecris.eurocris.org/handle/11366/135

Clements, A., McCutcheon, V., 2014. Research data meets research information management: Two case studies using (a) pure CERIF-CRIS and (b) EPrints repository platform with CERIF extensions. In: *CRIS2014: 12th International Conference on Current Research Information Systems*, Rome, May 13-15, 2014. http://dspacecris.eurocris.org/handle/11366/184

Clements, A., Proven, J., 2015. The emerging role of institutional CRIS in facilitating open scholarship. In: *LIBER Annual Conference 2015*, London, June 25th, 2015. http://dspacecris.eurocris.org/handle/11366/393

CODATA-ICSTI Task Group on Data Citation Standards and Practices, 2013. *Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data*. Report, CODATA, Paris. http://dx.doi.org/10.2481/dsj.OSOM13-043

Davidson, J., Molloy, L., Jones, S., Kejser, U. B., 2014. Emerging good practice in managing research data and research information within UK universities. In: *CRIS2014: 12th International Conference on Current Research Information Systems*, Rome, May 13-15, 2014. http://dspacecris.eurocris.org/handle/11366/201

Doorn, P., 2014. Going Dutch: Aggregating research information and research data services. In: *euroCRIS Strategic Membership Meeting Autumn 2014*, KNAW, Amsterdam, Nov 11-12, 2014. http://dspacecris.eurocris.org/handle/11366/355

Duchesne, S., Garcia, G., 2014. beQuali : une archive qualitative au service des sciences sociales. In: Cornu, M., Fromageau, J., Müller, B. (Eds.), *Les archives de la recherche. Problèmes et enjeux de la construction du savoir scientifique*. L'Harmattan, Paris, pp. 35-56. https://hal.archives-ouvertes.fr/halshs-00922690/

Duke, M., Ball, A., 2012. How to cite datasets and link to publications: A report of the digital curation centre. In: *23rd International CODATA Conference*, 2012-10-27 - 2012-10-31, Taipei. http://opus.bath.ac.uk/32421/

Elbæk, M. K., Sandfær, M., Simons, E., 2010. CRIS/OAR interoperability workshop. In: *CRIS 2010: Connecting Science with Society - The Role of Research Information in a Knowledge-Based Society. 10th International Conference on Current Research Information Systems*, June 2-5, 2010, Aalborg, Denmark. http://www.eurocris.org/Uploads/Web\%20pages/cris2010_papers/PPT/CRIS_OAR_Aalborg_4june2010_ms-kopi.pdf

Ensom, T., Wolton, A., 2012. *RDE metadata profile for EPrints*. UK Data Archive, University of Essex. http://www.data-archive.ac.uk/media/375386/rde_eprints_metadataprofile.pdf

Ginty, K., Kerridge, S., Cranner, P., McCutcheon, V., Clements, A., 2012. C4D (CERIF for datasets) – an overview. In: *CRIS2012: 11th International Conference on Current Research Information Systems,* Prague, June 6-9, 2012. http://dspacecris.eurocris.org/handle/11366/122

Hodson, S., 2013. New requirements for dataset metadata: a perspective from CODATA. In: *11th euroCRIS Strategic Seminar: "Metadata in Research Information Systems"*, Brussels, Sep 9-10, 2013. http://dspacecris.eurocris.org/handle/11366/278

Hogenaar, A., van Meel, M., Dijk, E., 2010. What are your information needs? Three user studies about research information in the Netherlands, with an emphasis on the NARCIS portal. In: *Publishing in the networked world: Transforming the Nature of Communication, 14th International Conference on Electronic Publishing, ELPUB*. Helsinki, Finland, 2010. pp. 290-303. http://elpub.architexturez.net/doc/oai-elpub-id-120-elpub2010

Jeffery, K. G., 2004. The new technologies: can CRISs benefit? In: *CRIS2004: 7th International Conference on Current Research Information Systems*, Antwerp, May 13-15, 2004. http://dspacecris.eurocris.org/handle/11366/311

Jeffery, K. G., 2012. CRIS in 2020. In: *CRIS2012: 11th International Conference on Current Research Information Systems*, Prague, June 6-9, 2012. http://dspacecris.eurocris.org/handle/11366/119

Koskela, R., 2013. Dataset metadata. In: *11th euroCRIS Strategic Seminar: "Metadata in Research Information Systems"*, Brussels, Sep 9-10, 2013. http://dspacecris.eurocris.org/handle/11366/267

Lewis, J. A., 2014. *Research data management technical infrastructure: A review of options for development at the University of Sheffield*. Report, University of Sheffield. http://figshare.com/articles/A_Review_of_Options_for_the_Development_of_Research_Data_Management_Technical_Infrastructure_at_the_University_of_Sheffield/1092561

Littauer, R., Ram, K., Ludäscher, B., Michener, W., Koskela, R., 2012. Trends in use of scientific workflows: Insights from a public repository and recommendations for best practice. *International Journal of Digital Curation* 7 (2), 92-100. http://dx.doi.org/10.2218/ijdc.v7i2.232

Luzi, D., Di Cesare, R., Ruggieri, R., 2012. Toward the integration of datasets in the CRIS environment: A preliminary analysis. In: *CRIS2012: 11th International Conference on Current Research Information Systems*, Prague, June 6-9, 2012. http://dspacecris.eurocris.org/handle/11366/104

Lyon, L., Pink, C., 2012. *University of Bath Roadmap for EPSRC. Compliance with research data management expectations*. Report, University of Bath. http://opus.bath.ac.uk/31279/

Matthews, B., Wilson, M. D., Kleese Van Dam, K., 2002. Accessing the outputs of scientific projects. In: *CRIS2002: 6th International Conference on Current Research Information Systems*, Kassel, August 29-31, 2002. http://dspacecris.eurocris.org/handle/11366/149

McGrath, A., Cox, M., 2014. Research excellence and evaluation using a CRIS: a cross-institutional perspective. In: *CRIS 2014*, *12th International Conference on Current Research Information Systems*, Rome, May 13-15. http://dspacecris.eurocris.org/handle/11366/198

Merceur, F., 2015. SEANOE : Publiez et citez vos données marines! In: *PREDONx Workshop on Scientific Data Preservation*, Strasbourg, December 9, 2015. https://indico.cern.ch/event/454764

OECD, 2006. Recommendation of the Council concerning access to research data from public funding. 14 December 2006 - C(2006)184, Organisation for Economic Co-Operation and Development, Paris. http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=159

Pierrel, J.-P., 2014. Ortolang: Une infrastructure de mutualisation de ressources linguistiques écrites et orales. *Recherches en Didactique des Langues et Cultures* 11 (1), 169-190. https://hal.archives-ouvertes.fr/hal-01109520

Príncipe, P., Rettberg, N., Rodrigues, E., Elbæk, M. K., Schirrwagen, J., Houssos, N., Jörg, B., 2014. OpenAIRE guidelines: supporting interoperability for literature repositories, data archives and CRIS. In: *CRIS2014: 12th International Conference on Current Research Information Systems*, Rome, May 13-15, 2014. http://dspacecris.eurocris.org/handle/11366/231

Prost, H., Malleret, C., Schöpfel, J., 2015. Hidden treasures. Opening data in PhD dissertations in social sciences and humanities. *Journal of Librarianship and Scholarly Communication* 3 (2), eP1230+. http://dx.doi.org/10.7710/2162-3309.1230

Rebouillat, V., 2015. *Archives ouvertes de la connaissance : Valoriser et diffuser les données de recherche*. Master's thesis, ENSSIB, Villeurbanne.

Ribeiro, C., 2013. UPBox and DataNotes: a collaborative data management environment for the long tail of research data. In: *euroCRIS Membership Meeting Autumn* 2013 (Universidade do Porto, Nov 14-15, 2013). http://dspacecris.eurocris.org/handle/11366/59

Ribeiro, L. M., de Castro, P., Mennielli, M., 2015. Surveying CRIS and IR across Europe. In: *EUNIS 2015, The Journey to Discovery.* The Abertay University, 10-12 June 2015, Dundee, Scotland, UK. http://fr.slideshare.net/LgiaMariaRibeiro/surveying-cri-ss-and-irs-across-europe-eunis15

Ribeiro, L., de Castro, P., Mennielli, M., 2016. *EUNIS – EUROCRIS joint survey on CRIS and IR*. Final Report, ERAI EUNIS Research and Analysis Initiative, Paris. http://www.eunis.org/wp-content/uploads/2016/03/cris-report-ED.pdf

Schöpfel, J., Prost, H., 2016 (forthcoming). Research data management in social sciences and humanities: A survey at the University of Lille 3 (France). *LIBREAS. Library Ideas* 29.

The Royal Society, 2012. *Science as an open enterprise. Summary report.* The Royal Society Science Policy Centre, London. https://royalsociety.org/\~{}/media/policy/projects/sape/2012-06-20-saoe-summary.pdf

Vanhaverbeke, H., Beullens, S., Timmermans, L., Gijsbers, K., Bras, B., Maeyaert, C., 2014. Deceiving simplicity. Balancing the need for ready-to-use research information with the semantic and technical complexity of research data. In: *CRIS2014: 12th International Conference on Current Research Information Systems*, Rome, May 13-15, 2014. http://dspacecris.eurocris.org/handle/11366/187

Wilkinson, M. D. et al., 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3, 160018+. http://dx.doi.org/10.1038/sdata.2016.18

On the authors

Joachim Schöpfel is Lecturer of Library and Information Sciences at the University of Lille 3 (France), Director of the French Digitization Centre for PhD theses (ANRT) and member of the GERiiCO research laboratory. He was Manager of the INIST (CNRS) scientific library from 1999 to 2008. He teaches Library Marketing, Auditing, Intellectual Property and Information Science. His research

interests are scientific information and communication, especially open access, research data and grey literature. He is member of euroCRIS.

Hélène Prost is an information professional at the Institute of Scientific and Technical Information (CNRS) and associate member of the GERiiCO research laboratory (University of Lille 3). She is interested in empirical library and information sciences and statistical data analysis. She participates in research projects on evaluation of collections, document delivery, usage analysis, grey literature and open access, and she is the author of several publications.

Violaine Rebouillat is Ph.D. student at the National Conservatory of Arts and Crafts (Cnam). She holds a Master's degree in information sciences and is interested in research data, especially in which way(s) a data culture emerges in a research team. She is also involved in a mission of the Digital Scientific Library (BSN), which consists in identifying the existing data management services in France.