Current research information systems and institutional repositories: From data ingestion to convergence and merger

Joachim Schöpfel, Otmane Azeroual

Current Research Information Systems and Institutional Repositories: From Data Ingestion to Convergence and Merger

Abstract

Current research information systems (CRIS) and institutional repositories (IR) were developed as clearly distinguished systems, with different objectives and functionalities, with different standards and data models, and for different needs and user groups. While academic librarians are often deeply committed to the management of open access and IR, they are less involved and familiar with research evaluation and CRIS. After a period of separate implementation of CRIS and IR, both systems started to converge and even to merge. Today IR often fulfil requirements of monitoring and assessment of institutional research performance while CRIS, beyond processing of metadata, begin to store, preserve and disseminate research papers. This chapter describes and explains the underlying dynamics, with examples to illustrate the benefits, the risks and potential barriers of this convergence. Special attention is paid to the role of data quality and user acceptance, and to the implications for the academic librarian.

Keywords

Current research information system, institutional repository, research information management, data quality, user acceptance, academic library, open access, open science

Authors

Joachim Schöpfel
University of Lille, GERiiCO Laboratory (France)

Otmane Azeroual
German Institute for Higher Education Research and Science Studies (DZHW), Berlin (Germany)

Introduction

Current research information systems (CRIS) and institutional repositories (IR) were developed as clearly distinguished systems, with different objectives and functionalities, with different standards and data models, and for different needs and user groups. IR have been described as ‘‘a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members (...) including long-term preservation where appropriate, as well as organization and access or distribution’’ (Lynch 2003). CRIS, on the other hand, stand for a family of information systems (databases, decision support systems etc) to store and manage data about research conducted at an institution (Joint 2008). While academic librarians are often deeply committed to the management of open access and IR, they are less involved and familiar with research evaluation and CRIS.

Both systems are part of the ‘‘fourth paradigm’’ (Hey et al. 2009), data-intensive scientific discovery, a new way of doing science based on information and communication technology. After a period of an *‘alleged dichotomy’’* (de Castro et al. 2014), of separate implementation and of discussion on data ingestion, data exchange and interoperability, both systems started to converge and even to merge.

Today IR often fulfil requirements of monitoring and assessment of institutional research performance while CRIS, beyond processing of metadata, begin to process, store and disseminate research papers.

Our chapter tries to make this evolution understandable for academic librarians, based on an overview on recent papers and surveys and with examples to illustrate the challenge for the library. What are the benefits and what are the risks and potential barriers of this convergence? Special attention will be paid to the role of data quality and of user acceptance, and to the implications for the academic librarian.

Institutional repositories

Institutional repositories (IR) are part of the green road to open access (OA), through author self-archiving of their papers in university-driven online collections or databases (Harnad et al. 2004). Being part of the OA movement, they are a new vector of scientific information which ‘‘have no counterpart in the traditional landscape of scholarly communication’’ (Suber 2012, p.52). OpenDOAR [1], the global directory of OA repositories lists close to 4,000 IR in 2019. There are other kinds of open repositories, governmental, disciplinary or aggregating. Yet IR are the most important category, representing 87% of all open repositories. Many are from European institutions; others are from the Americas and other parts of the world. Many of the highest-ranked universities have their own IR, like Harvard, Columbia, MIT, Oxford or Cambridge.

Often, IR have been implemented as standalone systems, as a platform on an institutional server more-orless connected with other systems. Some IR have been launched with a regional or national infrastructure, for instance as a portal on a governmental open repository. Many IR are based on open software, like DSpace, EPrints, Islandora or WEKO; some of them are designed for a specific national environment, like OPUS or MyCoRe in Germany or HAL in France [2]. Other IR are launched with commercial software, like Digital Commons, or are in-house developments.

The initial and main functions of IR are direct communication, i.e. immediate and open dissemination, and preservation of the institutions’ scientific output, which means the papers and data produced by their affiliated researchers. Generally, IR are initiated in compliance with the goals of the global OA movement, with the purpose of increasing the accessibility and findability of scientific knowledge, for other researchers, for the industry and for society as a whole. Also, most IR have implemented a standard metadata format, compliant with the Dublin Core, and a standard protocol for harvesting and other aggregating services, and especially the OAI metadata harvesting protocol.

Academic librarians are most often (if not always) deeply involved and committed to IR, from the very start of the project on to the administration, maintenance and development of the system, to content updates, metadata curation and staff training. Now, research managers and policy makers in particular ask for indicators of institutional research performance in order to assess scientific excellence and international competitiveness; also, they identify IR as potentially relevant data sources for in-house statistics and indicators. Over the years, IR adapted to the requirements of monitoring and assessment of institutions’ scientific output, with new functionalities and services. This new (secondary) function impacts the IR twice. Firstly, monitoring requires metadata, not the deposit of text or data files. Therefore, for instance, the number of documents in the French HAL repository declined over time to only 33% of the total because research structures increasingly upload metadata without deposits, for needs of evaluation. Secondly, as monitoring requires reliable and comparable information sources, this new function fosters standardization of metadata formats, of persistent identifiers like DOI and ORCID for resources, authors, institutions and so on, and of terminology (nomenclature) for disciplines,

subjects, methodology, countries, towns and other relevant categories. IR are often hosted and run by academic libraries; the fact that academic librarians are used to norms and standards contributed to this process in a helpful way.

Initially designed for articles and preprints, IR contain other resources, like theses and dissertations, conference papers, unpublished reports or working papers; increasingly IR also contain courseware, software, image files, AV materials and datasets. However, we must be careful with the meaning of ‘‘contain’’; more and more, IR do not only accept author self-archiving of papers and other items but allow the creation of metadata records without the document or data files, for the reason mentioned above. Therefore, many IR are today a kind of ‘‘mix’’ of OA repositories and bibliographic databases, with a more or less important section of freely accessible resources, along with metadata records linking to other platforms (articles) or repositories (data) via persistent identifiers or URLs.

Current research information systems

Research information includes all metadata that arises in connection with research activities, such as information about publications, project data and persons involved. Because this information is often stored in multiple systems, current research information systems (CRIS) are needed to bundle the information in a structured way, to simplify the creation of reports and to enable value-added services. Academic institutions can be supported by the provision of CRIS (see Fig.1). A CRIS is a database or federated information system, representing research management information (Jeffery 2012). ‘‘The CRIS provides a single portal bringing corporate and academic research activity together, reducing duplicate data entry, increasing data quality, identifying authority sources of information and recording complex relationships between researchers, projects, outputs and impact’’ (Clements and Proven 2015). Relevant research information in a CRIS database is for instance data on people (name, job title, affiliation, skills…), organizations and research facilities (name, type, country…), projects (name, duration, funding, programme…) and outputs (publications, research data, patents…).
Research information should not be confused with research data (research output). With the help of CRIS, the entire research process in academic institutions can be supported and the research context documented. For example, CRIS should be able to link projects with funding and research results and allow for evaluation and assessment within an institution and comparison with other institutions. Furthermore, a CRIS can be used to manage research projects, research results, research resources and research funding.

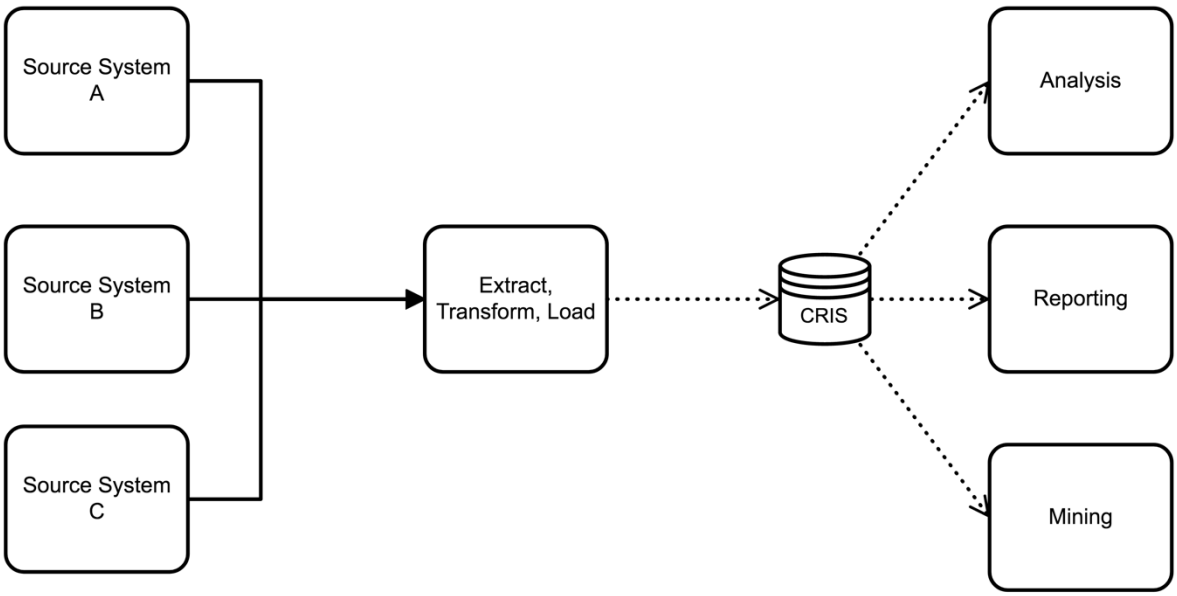

Fig.1. Architecture of a CRIS.

CRIS are facing a certain challenge, as for example divergent user expectations. From the perspective of administration of a CRIS, research information should be presented as uniformly as possible in order to be able to make comparisons and assessments quickly and easily. CRIS, as a central database of all research relevant high-quality information of an institution, make it possible for prospective customers and cooperation partners to easily find the responsible persons and contact information as well as provide an overview of the research activities and services. Researchers, on the other hand, need a precise description of their research profile, taking into account specific criteria for evaluating the activity of the respective discipline, or they want to find exact information and use the subject-specific research information of other researchers. CRIS can also support auxiliary functions for researchers: for example, the creation of CVs and publication lists, or other re-use aggregated data from CRIS on a website.

Another challenge facing CRIS, from an international perspective, is the compatibility and interoperability of research information coming from different systems. In Europe, USA, China and other countries, the trend towards the emergence of international CRIS can be observed, which requires standardization in the field of research information. Standardization is key, and different initiatives support the standardization of CRIS, like the Common European Research Information Format (CERIF), the Consortia Advancing Standards in Research Administration Information (CASRAI) and the German Research Core Dataset (RCD/KDSF).

Academic librarians are generally involved in CRIS projects but in a different way, unlike their involvement in IR. They are less involved in development and administration and more involved in standardization issues, data curation and interconnection with catalogues and repositories. Nonetheless, they are key stakeholders in the success of any CRIS project.

Convergence

In his short history of research information management, Keith Jeffery highlights that compared to IR, CRIS have a longer history, going back more than 50 years (Jeffery 2012). The two software categories have not been developed together, at the same time, in the same technological environment. When

institutions started to launch open repositories, CRIS – at least some kind of research information software - often were already there. The first impact was, therefore, that CRIS discovered IR as a new source of potentially relevant information.

Data from IR to CRIS

A CRIS needs relevant, reliable and valid information on scientific publishing as one part of research output (results). For instance, it needs information about the type and status of publication, the date and the publisher, and it needs information about the relationship between this output and persons (authors), projects, organizations (affiliations) and other outputs (patents, data and so on). Insofar as this kind of information is stored in institutional repositories, CRIS managers have identified IR as a relevant data source for CRIS, as one of the ‘‘source systems’’ in Figure 1, similar to bibliographic databases and library catalogues.

Figure 2 shows the example of the dataflows of the ZORA repository of the University of Zurich [3]. The dataflow from IR to CRIS is on the right-hand site, from ZORA to the research information system AKABER, shown in blue. The content of the flow is bibliographic data. The purpose is academic reporting. Because of this dataflow, the AKABER system has a near 100% coverage of the academic output, with library curated data and a high proportion of papers in open access.

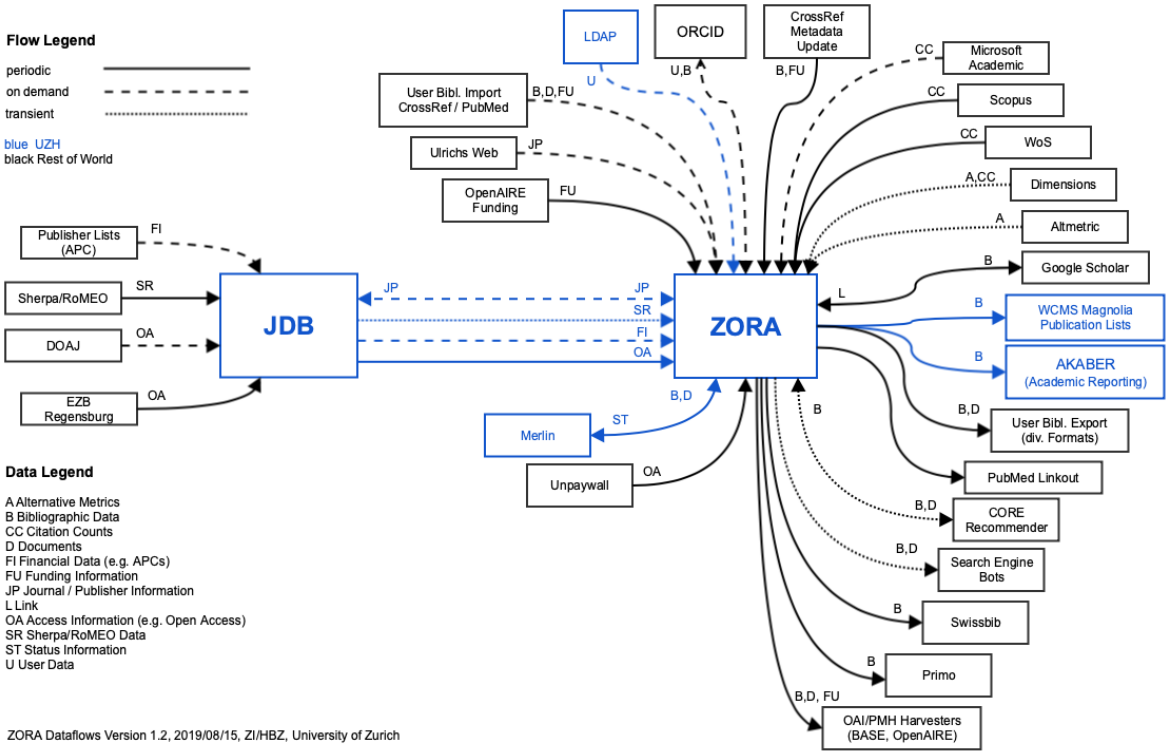

Figure 2. Dataflows of the ZORA repository, University of Zurich (source: Martin Brändle, UZH)

In the words of Keith Jeffery, a CRIS is a kind of gateway to knowing where all the relevant resources are (2012). For instance, when a new CRIS was implemented at the Free University of Bozen-Bolzano, it ingested data from four information systems: human resources, purchasing, the national Italian publication database from Cineca and the local IR; the CRIS maps the attributes of the uploaded data and saves them into the corresponding CRIS tables. Yet, as the CRIS was designed as the unique data

entry point for the researchers' publication data, it also transferred data back to the IR (Siciliano et al. 2014).

Data from CRIS to IR

On the other side, institutions, and especially academic libraries, identified CRIS as a source of valid information on research output (articles, books, dissertations and conference papers) to improve the quality of metadata (missing or erroneous data) and of the coverage (missing records). IR data is not always consistent, especially because of self-archiving, and the usually high degree of standardization makes CRIS potentially interesting for the improvement of the IR quality. CRIS, in the words of Joint (2008), ''have given a great boost to open access repositories''.

The Norwegian national CRIS has been developed as a ''single point of entry for all research'' (Wenaas et al. 2012). Authors are invited to upload their publications on the CRIS platform; but as the CRIS also receives metadata from databases and journal platforms, it can enrich the local IR and inform institutions and authors about potential uploads (legal deposits allowed by publishers), which means that ''there is a potential to increase holdings in the IR and also identify the large portion of Norwegian research results that are candidates for self-archiving''. The article submission form on the CRIS platform is easy and user-friendly, and it is the starting point for an invisible ''highway'' to the IR.

Ivanovic et al. (2014) present plugins for the export of data on published results, in particular PhD dissertations, and conferences via standardized (OAI-PMH, SRU) and non-standardized XML protocols and in different formats (Dublin Core, MARC21, ETD-MS) to various systems, including their local IR but also a national government gateway (Serbia), OpenAIRE (European projects) and DART-Europe (European dissertations). The key is standardized metadata and common formats and protocols, including shared and centralized identifiers. Here, a local CRIS is feeding local, national and international systems with new data about scientific output.

The OpenAIRE gateway to European research output illustrates this. OpenAIRE was launched in 2009 as a follow-up project of the DRIVER infrastructure and aggregator of European IR [4]. Since 2015, OpenAIRE provides guidelines for CRIS managers ''to expose their metadata in a way that is compatible with the OpenAIRE infrastructure (i.e. to support) the inclusion and therefore the reuse of metadata in their systems within the OpenAIRE infrastructure'' which is described as being 'itself a CRIS system' [5].

Along with new bibliographic information for updates, CRIS also contain useful information to enrich or disambiguate existing IR metadata, especially about the authors and institutions (for example spelling, identifier or address), but also about projects (project name, funding body or project tender).

Impact on both systems

A recent survey from EUNIS and euroCRIS, with 84 institutions from 20 European countries, shows that 65% of institutions have linked their CRIS and IR (Ribeiro et al. 2016). The communication between IR and CRIS and their gradual convergence impact both systems, but in different ways. On the one hand, CRIS must take into account and adjust to library and open access standards. Clements & Proven (2015) describes the evolution of an institutional CRIS ''from [its] traditional role as a tool managed by the Research Office to manage and assess research towards more widespread uses within institutions, in particular within the Library, to facilitate Open Science''. For this reason, a CRIS is increasingly the primary bibliographic record of the scientific output and must be compliant with the usual

bibliographic standards and procedures. Clements and Proven (2015) observe that a similar role is emerging for research data and that CRIS become the primary source of research data records, which means that CRIS must also be compliant with data repository standards, such as generic and discipline-specific metadata formats for research data.

This convergence also impacts the IR. The first and principal impact is on the workflows and the functions of IR, the gradual replacement of self-deposits in the IR by data (document files) and metadata (records) imported out of the CRIS, from different origins (self-deposits in the CRIS, library catalogues, metadata from publishers), contributing to a ''complex two-way exchange of data'' between CRIS and IR (MacGregor 2019). Beyond this fundamental change of workflows and functions, the interconnection with a CRIS may have other effects on the local IR. ''Progressively, repository architectures are merging and co-evolving with CERIF-CRIS'' (Jeffery 2012). Siciliano et al. (2014) describe how the integration of both systems can require a new and much more detailed hierarchical organization of collections and communities in the repository to allow a consistent mapping with the CRIS. More generally, as the purpose of the standard CRIS format CERIF is to represent research information and to transfer it between systems, it can improve the interoperability and compliance with the FAIR principles, of repositories and other digital archives (Engelmann et al. 2018).

From local to regional or national systems

If CRIS and IR converge first and foremost on a local level, within the framework of the institutional architecture of information systems, the process can also be observed on a larger, regional or national level. We already mentioned the case of the national CRIS in Norway and the national gateway to publications in Serbia. In the Netherlands, the Dutch national portal of research information NARCIS [6] aggregates data and metadata from both repositories and research information systems (Dijk et al. 2006). Based on information imported from local CRIS and IR, NARCIS provides information on research institutes (profiles, addresses, projects, and publications), researchers (expertise, addresses, research projects, and publications), research activities (research projects and programmes), publications (metadata, and full text), datasets (metadata) and news (web pages). The main problem of NARCIS is data quality, especially the matching of the researchers' names from different sources (library catalogues, CRIS, IR), and for this reason the Netherlands have developed a national digital author identification (DAI), for all systems and levels.

The NARCIS infrastructure was the model for the launch of a similar research portal in Catalonia, PRC [7]. The PRC portal receives information from the local CRIS of Catalan universities and research organizations. It is based on DSpace-CRIS but it does not include documents, just metadata. Here, the convergence remains on the local level, at least for the moment.

Merger

Clearly, there are ''overlapping areas in the tasks that (both systems) perform and there has been a gradual convergence of these two types of systems'', through systematic metadata transfer or ''by allowing one of them to take over the features of the other one, thus delivering a single-system integrated functionality'' (de Castro et al. 2014). This opens the door for further convergence to a gradual merging of both systems, not only at the user interface but also in the back office. The EUNIS/euroCRIS survey reveals ''CRIS acting as repositories, repositories with extended data models, a wide range of interoperability features between co-existing CRISs and repositories and even a new

species in the ecosystem that claims to be both a repository and a CRIS'' (Ribeiro et al. 2016). This can be described in two ways.

CRIS with IR functionalities

First, we can observe how CRIS have developed functionalities usually attributed to IR; ''CRIS acting as repositories'', in some way or other. For instance, some CRIS allow the deposit of documents, a service which is a basic feature of open repositories. Labastida i Juan (2015) describes the bridge between the IR (DSpace) and the in-house developed CRIS at the University of Barcelona (Catalonia). Authenticated researchers self-archive their publications not in the IR but in the CRIS. After validation, the CRIS transfers the documents in PDF with standardized metadata (Metadata Encoding and Transmission Standard) to the IR which attributes a persistent identifier (handle) to each deposit and communicates this identifier to the CRIS. The CRIS also controls embargo periods and re-use rights (licensing). Interoperability through standardization is the key to making the two operating systems communicate with each other. Since 2015, the quality of the transferred information has been improved, for instance by linking projects with outputs in order to get the project metadata, including a controlled thesaurus for subjects or trying to get a better control of authorities; however, IR and CRIS remain two distinct systems (personal information by Labastida i Juan, October 2019).

One of the leading commercial research information management solutions, Pure, from Elsevier, is perhaps the best example of this evolution. Initially developed by the Danish start-up Atira as a ''normal'' CRIS, with all required functionalities such as evaluation and reporting, identification and dissemination of expertise and interconnection with national assessment frameworks. Pure moves progressively towards a ''single content management system on campus'', explicitly labelled as a repository [8]. Pure offers the possibility to ingest and store not only the data on research but also the related content and documents -, the scientific output itself - such as publications, datasets and awards, press clippings, ''narrative data around projects'' and so forth. ''With Pure, you can store your institution's manuscripts, data sets and other research artifacts in a single place where they are linked to your researchers' profiles and other research data''. Thus, central IR functions have been transferred to the CRIS, upstream of the workflow, leaving to the IR the preservation and dissemination of the documents and records, along with the attribution of a persistent document identifier. The researchers interact with the CRIS for deposit, creation of metadata and evaluation, while the IR remains available for the retrieval of papers (Figure 3).

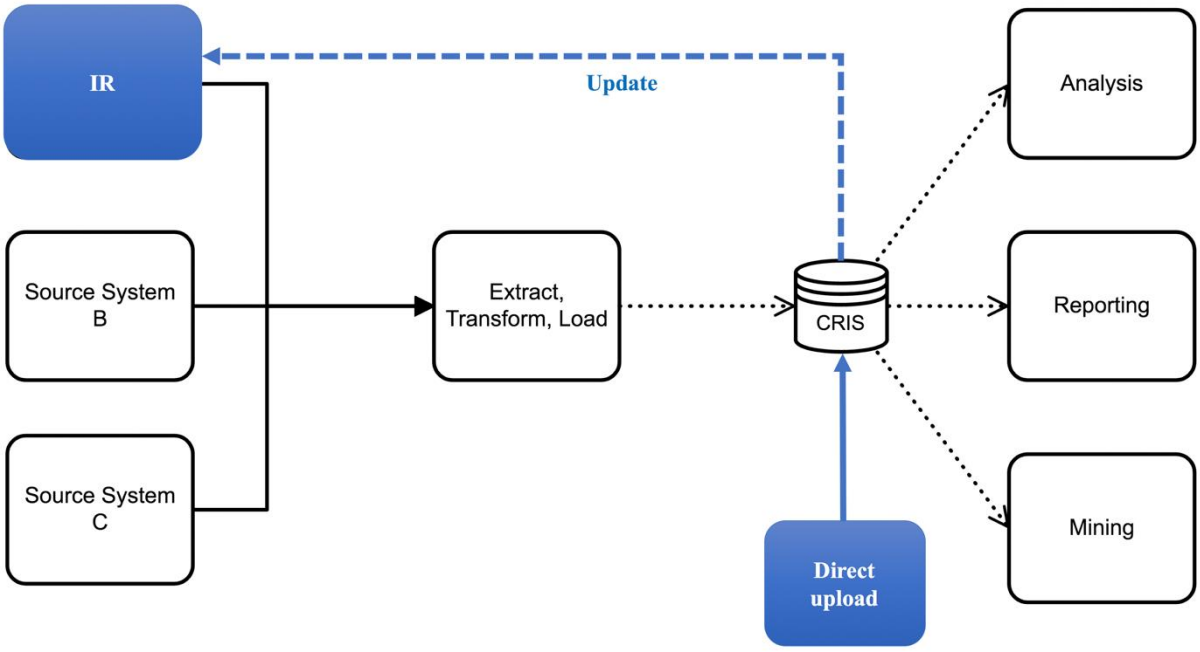

Figure 3. CRIS with IR functionalities

Some CRIS provide the required conditions for the dissemination (harvesting) of data and metadata via the OAI-PMH protocol and/or import metadata and files from commercial databases, like Scopus, with a focus on journal publishing. Thus, they can play the role of an operational IR, from the ingestion of content to its dissemination. Sometimes, ‘‘an independent IR (remains) in operation for disseminating other kinds of institutional output such as dissertations and/or grey literature’’ (de Castro 2014). But the potential for a kind of vertical integration between CRIS and IR on the local level is real. The challenge is organizational than technical: repositories are mostly hosted by academic libraries while CRIS are not. Another challenge is shared with many IR: how can CRIS guarantee long term preservation of deposited contents?

IR with CRIS functionalities

Second, we can observe how open repositories begin to develop functionalities and services usually attributed to research information management, such as scientometric assessment of the scientific production of individual researchers, institutions and departments, the generation of individual CV, the linking between funded research projects, publications and patents, and so on.

The Polish Omega-Spir software is one example. Initially developed as an IR, from 2013 on its functionality ‘‘went beyond the typical functionality of institutional repository towards the functionality of CRIS … Due to applied intelligent tools (acquisition tools, reporting functionalities), the maintenance efforts of Knowledge Base are essentially reduced compared to the typical solutions. The process of moving the system to the University level was already simpler, as the team had experience with organizational and training issues at the faculty level’’ (Rybinski et al. 2018). The authors mention two major benefits, no costs in excess of usual IR maintenance, and a single unique interface. A third benefit should be added: complete institutional control over the whole process and content, without outsourcing.

The successful extension of DSpace, the free open source software for repositories for research information management is another example of this merger. ‘‘Differently from other [commercial] CRIS..., DSpace-CRIS has the institutional repository as its core component, providing high visibility on the web to all the collected information and objects’ [9]. The flexible data model of DSpace-CRIS is compliant with the CERIF standard and the ORCID identifier and allows for the configuration of different data models and metadata schemas, according to local requirements. The DSpace-CRIS wiki indicates more than 100 installations worldwide, mainly in Europe (Italy, Germany, Spain) and Asia.

HAL is a third example. Initially launched as an open repository based on self-archiving of research papers by French academic communities, HAL developed into a platform with a high number of institutional portals and other collections. Answering the institutional need for scientific monitoring and evaluation, HAL adds features and plugins useful for scientometrics and the assessment of research output. This development has two main effects on HAL: the increasing use of national and international standards and identifiers (nomenclature of publication types, DOI, ORCID), and the number of records without the document files (though only 24% of the HAL metadata provide access to the document).

Data quality

Research information is often spread across different systems within an institution. This results in an increased administrative overhead, inconsistent and incomplete data, and unreliable reports that negatively impact business decisions. Therefore, the users of the CRIS require high quality data. Increasing data quality is an important and still problematic task for many organizations. Although effective quality management should occur as early as possible in the data flow (as for example in manual data collection), quality deficiencies often only become visible in the CRIS (Azeroual and Schöpfel 2019). The data quality can be measured based on certain criteria such as completeness, correctness, consistency or timeliness (Azeroual et al. 2018b) and can be positively influenced during the phase of data acquisition and integration, for example, when transferring research information from various data sources into the CRIS. A proactive data quality management means not only the removal of random errors but also:

- the identification and elimination of sources of error,
- a constant control of the research information and its quality,
- preventive measures (such as pro-active measures) to prevent further errors and
- the regular cleanup of new data errors (for example, using data cleansing).

It is not enough to clean up the research information only when it is integrated into the CRIS, but it is necessary to communicate errors to the appropriate places so that they are already remedied in the source system, such as IR, catalogues, human resources, accounting. This not only prevents the occurrence of the same errors the next time the data is loaded from the respective system into the CRIS by Extract, Transform, Load (ETL) process (Azeroual et al. 2019a), but also sensitizes the responsible persons, for example, when manually entering research information.

Whether research information can achieve high quality in CRIS depends on the quality of the respective data sources. It is easy to overlook the question of whether a data source is trustworthy at all and whether it has a high-quality database from the outset (Azeroual et al. 2019b). However, if the data quality is taken seriously, one should also rate the data sources according to their trustworthiness. In order to be able to analyze research information, that information must first be sensibly selected with regard to the intended use. The amount of data selected should be as large as necessary and as small as possible. The right choice of research information is a first step towards high data quality. A general

statement about which research information should be selected for integration cannot be made. This depends on the individual needs of the particular organization.

The problem with IR is that they are often not exhaustive and that they may contain incomplete or erroneous records. This may not be a crucial issue for IR insofar as their goal is dissemination, and therefore recall sensitivity (retrieving most of the relevant content) seems more important than precision. However, this is unacceptable for CRIS where precision of information retrieval is essential. Eliminating erroneous research information when integrating it into a CRIS is often a cumbersome but important process for creating high data quality. Therefore, a workflow of data cleansing should be provided for organizations so that they can correct the occurring quality problems and ensure the quality in their system (Fig. 4). This has advantages for the organizations, both internally and externally. Establishing a data quality model is sustainable, which means quality ensures the quality of service and knowledge available in organizations. It helps to bundle important know-how and keep it in organizations, independently of people.

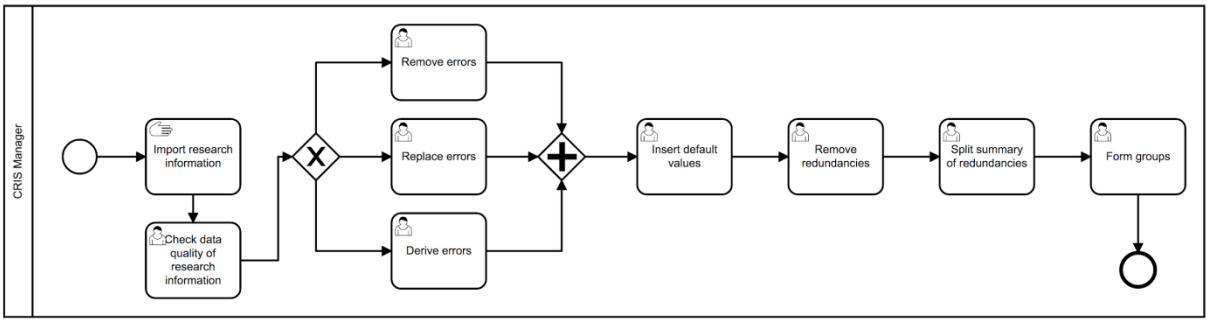

Fig.4. Workflow of data cleansing.

The key elements of this workflow can be described as follows:

(1) For records that cannot be corrected from the outset, their preliminary deletion in the integration process should be considered. For example, this may be the case of missing or incorrect key attributes or other unique identifying attributes such as part numbers and descriptions. Such a master data record may not be clearly identified without this information and has no value in the CRIS. Such records must first be corrected or re-entered in the respective department.

(2) Missing or incorrect data records can be corrected by publication data. For example, if there is only a unique article number without an article description, it is usually possible to determine the associated article description from it and insert it into the attribute provided for this purpose.

(3) Similar to the replacement of research information from publication databases, certain values in data integration can be derived from calculations or other logic for missing or erroneous values.

(4) Using meaningful default values instead of missing data can prevent null values and improve data quality. Examples of this would be date fields, which by default are set to the current date or any other specified day when not filled. It is important to define the default values in advance in such a way that all possible cases and exceptions are recorded and no further errors result from incorrect insertion logic (which are in hindsight difficult to identify).

(5) Any redundancies that have arisen or been identified as part of the integration of research information should be treated as part of the data cleansing. A good way to do this is to consolidate (summarize) multiple records into a single correct one. It is always important to

find a balance between effectiveness (finding as many redundancies as possible) and efficiency (keeping runtime as short as possible).

(6) If records are composed of non-related attributes, they must be separated and stored in two different records. Such erroneous data sets can occur, for example, in the automated summary of redundancies. A manual check or automated comparison with correct comparison values should be done in such cases.

(7) After the research information has been corrected, as a last step it should be checked to see whether there are already groups for the newly added data records in the CRIS into which they can be added. If not, check whether the new records form a separate group. Example of this are articles or author classes, which do not exist yet and are recognized only during the data integration. The modeled workflow examines both cases with the involvement of the respective departments.

Data cleansing activities are limited to relatively simple but sometimes time-consuming activities. Also, instead of improving the quality of the data, further errors may occur if, for example, data is automatically changed and there is insufficient manual control of the changes subsequently. Therefore, an iterative process is recommended during the cleanup. To ensure that only high-quality data is sent to a CRIS, a data cleansing tool must be used at CRIS-using facilities. For more detailed information on how the data cleansing processes work using a practical example in the context of CRIS (such as syntax errors, standardization, matching, merging, and enrichment), see Azeroual et al. (2018a).

The rapid development of CRIS has changed the way research organizations capitalize on research information. Because of this high information dependency, the effective management of data quality and ensuring data integrity are of the highest priority. If data quality issues are not detected and validated early, the risk is poor research information. Data cleansing is the first step in the data preparation process. It identifies and corrects errors in a record to ensure that only high-quality data is transferred to CRIS. When research information comes from multiple sources, such as catalogues, databases and repositories, the need for cleansing data increases because the sources have redundant data or incompatible data formats. Data cleansing helps ensure data accuracy, so only high-quality data is available for analysis and decision making. Whenever the erroneous data are from an IR, the cleansing process will be helpful to increase the quality of the repository itself, in the case of systematic errors, lack of standardization etc. However, as MacGregor (2019) mentions, the problem can also be the lack of open standards and protocols, interoperability and support by a commercial, proprietary CRIS.

User acceptance

The acceptance of an information system generally refers to users' decision on whether they should buy or implement it and further use it in the long term, in the sense of active willingness and not only in the sense of reactive toleration (Arnold & Klee 2016). Over the past 30 years, many different models have been developed to describe and link people, systems and contextual factors with potential impact on the acceptance of information systems. The most influential, most tested and best operationalized approach is the Technology Acceptance Model (TAM) (Davis 1989).

The TAM explains the acceptance of the CRIS on the basis of cognitive factors. Cognitive factors are understood as perceived usefulness and perceived ease-of-use operability of the examined information system. The perceived usefulness refers to a user's assessment of how the use of a specific IT application improves the execution of work tasks within a specific organizational acceptance context. The perceived ease of use relates to the assessment of whether the use of the information system can be learned without difficulty and effort. The higher the benefits and ease of use of a CRIS,

the more likely it is for users to be willing to use the system. In the TAM, the interaction of the two factors - perceived usefulness and perceived ease of use - results in an intention on the part of the user to use the technology in question (behavioural intention to use) which can lead to a real, actual system use (Schöpfel et al. 2019).

Perceived usefulness and perceived ease of use, are both valid indicators for the acceptance and use of CRIS. Yet both factors are impacted by other independent variables such as culture, job position and function, data quality and system security (Azeroual et al. 2019c). Following different studies (Lucke 1995, Kollmann 1998, Venkatesh et al. 2007), in Western cultures the perceived usefulness seems to be more important in determining the intentions and actual use, while ease of use appears to be the key in non-Western cultures. Communication regarding the introduction of the CRIS, such as training manuals and direct contact with the CRIS developer, should therefore carry different messages, depending on the culture in which the CRIS is to be used.

As for the job positions and functions in the academic environment of CRIS and IR, at least five internal user groups must be distinguished: researchers (scholars), students, administrators, scientific bodies (councils and similar), and university managers responsible for research strategies. As Rybinski et al. (2018) observe, the needs and expectations of these groups vary significantly, and in many cases they may even be contradictory. And then there may be external stakeholders with other needs and requirements, such as entrepreneurs, funding bodies, government authorities.

Introducing a CRIS to people working with library holdings, digital libraries and repositories may be a challenge. ''In every community any change and innovation in consolidated procedures and workflows is considered with suspicion' (Siciliano et al. 2014); Rybinski et al. (2018) had to cope with ''scepticism from the researchers'', from the beginning of their CRIS project. Reluctance, suspicion and scepticism may result not only from fear of change, but also from issues of control and evaluation, personal data and ethics. Faced with a CRIS project, some academic librarians may experience helplessness, loss of control and lack of skills; also, they ''must come to grips with the role of repositories within the CRIS environment'' (Joint 2008). Siciliano et al. (2014) admit they initially had underestimated problems related to privacy and copyright and their impact on the validation process. Traditional library ethics include promotion of open access and transparency, non-discrimination, respect of privacy, confidentiality, commitment to neutrality and personal integrity [10] while CRIS have more to do with performance, competition, assessment, evaluation and management. This may create conflicts.

How do we deal with these potential barriers? Acceptance is the key, especially by researchers and librarians. A good partnership between librarians and IT staff is required, just as a trusting relationship between librarians and researchers is necessary. IR are generally much lighter and easy to configure platforms than CRIS (de Castro et al. 2014), which implies more need for acceptance of CRIS than IR.

The experience of immediate usefulness may contribute to acceptance. For instance, in the Polish CRIS project, researchers could, at an early stage, ''immediately observe their profiles, so that they gradually turned to be more and more keen to contribute to the process of the (knowledge) database maintenance'' (Rybinski et al. 2018). The Italian team tried another way to achieve a faster and easier acceptance of the new CRIS: ''we decided … to feed it with project and publication data produced [in the last 17 years] before committing it to the research community … we are confident that this valuable critical mass of research outcomes easily searchable and reusable in the system [will] improve its acceptance'' (Siciliano et al. 2014). This may mean additional work for the implementation team; but the implementation of a CRIS and the convergence of CRIS and IR will show other benefits, improved quality, reliability and reusability of metadata and less work, avoiding duplicated inputs in the two systems.

Future perspectives

According to the aforementioned EUNIS/euroCRIS survey, 18% of the responding institutions already use the same software application for both CRIS and IR (Ribeiro et al. 2016). The convergence of both systems is more than a marginal phenomenon and reflects a strategic choice made by an increasing number of universities and other research institutions. This evolution is likely to continue because of benefits such as reduced costs for implementation, hosting and maintenance, improved data quality and the users' general preference for simple procedures ('one input, many outputs'), unique gateways and single points of contact. The ongoing vertical integration of data services in the information industry (Chen et al. 2019) will foster this development.

This development may take different paths, at a different pace, in different regions. Castro et al. (2014) observe that there were some geographical areas in the world with a 'fairly well-established open access repository network, very little CRIS implementation and barely any mechanisms in place yet for ensuring CRIS/IR interoperability'. Perhaps, in these regions 'convergence' means that their repositories will develop new features that meet their research information management requirements; perhaps they will turn 'their eyes to open source CRIS when comprehensive research reporting starts to be perceived as an institutional need'. Then, convergence takes place.

For academic librarians, this convergence is a challenge insofar they are usually working with repositories and digital libraries while 'representing research management information, CRIS are an integral part of the ICT environment providing the context for the day-to-day work of the researcher – or research manager, innovator or the media' (Jeffery 2012). But this challenge can be an opportunity for academic librarians to develop new services and skills, to demonstrate their mastery of metadata, curation and standards, and to improve their relationships with academic communities and research management. They have little choice and must adapt to CRIS, as 'when it comes to interoperability with legacy systems such as Finance and HR, CRIS are the preferred system to link to because of the data and information contained in them; CRIS will become more important than IR' (Ribeiro et al. 2016). CRIS are here to stay.

For this reason, the real challenge for academic libraries is data quality and standards. They can build on a long tradition and practice with cataloguing and bibliographic standards and rules. They have a good knowledge of academic users' needs and behaviours. And they have experience with service levels and developments. In the words of one of the leading figures of the CRIS community, Keith Jeffery, the vision of the future development of CRIS 'requires a user model, a processing model and a data model providing consistent services underpinned by a resource model' (Jeffery 2012). The key is metadata, and metadata is part of academic librarians' core competencies. Their expertise with unique identifiers, standard protocols and data formats is required for the development of both systems, and they will seize the convergence and merger of IR and CRIS as an opportunity for their future development on campus, as a central part of the research service environment.

Notes

[1] Directory of Open Access Repositories https://v2.sherpa.ac.uk/opendoar/
[2] Hyper articles en ligne https://hal.archives-ouvertes.fr/
[3] Zurich Open Repository and Archive https://www.zora.uzh.ch/
[4] History of OpenAIRE https://www.openaire.eu/history
[5] OpenAIRE Guidelines for CRIS Managers 1.1 (2018) https://zenodo.org/record/2316420
[6] National Academic Research and Collaborations Information System https://www.narcis.nl/

[7] Portal de la Recerca de Catalunya https://portalrecerca.csuc.cat/?locale=en
[8] Elsevier Pure https://www.elsevier.com/solutions/pure/features
[9] DSpace-CRIS https://wiki.lyrasis.org/display/DSPACECRIS/DSpace-CRIS+Home
[10] IFLA Code of Ethics https://www.ifla.org/publications/node/11092

Bibliography

Arnold, C., Klee, C. (2016). Akzeptanz von Produktinnovationen: Eine Einführung, Springer Fachmedien Wiesbaden GmbH. https://doi.org/10.1007/978-3-658-11537-1

Azeroual, O., Saake, G. and Abuosba, M. (2018a). Data quality measures and data cleansing for research information systems, *Journal of Digital Information Management*, 16(1), 12–21. https://arxiv.org/abs/1901.06208

Azeroual, O., Saake G., Wastl, J. (2018b). Data measurement in research information systems: metrics for the evaluation of data quality, *Scientometrics* 115(3), 1271–1290. https://doi.org/10.1007/s11192-018-2735-5

Azeroual, O., Schöpfel, J. (2019). Quality Issues of CRIS Data: An Exploratory Investigation with Universities from Twelve Countries. *Publications*, 7(1), 14. https://doi.org/10.3390/publications7010014

Azeroual, O., Saake, G. and Abuosba, M. (2019a). ETL best practices for data quality checks in RIS databases. *Informatics*, 6(1), 10. https://doi.org/10.3390/informatics6010010

Azeroual, O., Saake G., Abuosba, M., Schöpfel, J. (2019b). Solving problems of research information heterogeneity during integration – using the European CERIF and German RCD standards as examples. *Information Services and Use*, 39(1-2), 105–122. https://doi.org/10.3233/ISU-180030

Azeroual, O., Saake, G., Abuosba, M., Schöpfel, J. (2019c). Quality of research information in RIS databases: A multidimensional approach. *International Conference on Business Information Systems*, BIS 2019, vol 353, 337–349. https://doi.org/10.1007/978-3-030-20485-3_26

Chen, G., Posada, A., & Chan, L. (2019). Vertical Integration in Academic Publishing. Connecting the Knowledge Commons - From Projects to Sustainable Infrastructure: The 22nd International Conference on Electronic Publishing–Revised Selected Papers, 15–40. Marseille, OpenEdition Press.

Clements, A., & Proven, J. (2015). The emerging role of Institutional CRIS in facilitating Open Scholarship. LIBER Annual Conference 2015, London, June 25th, 2015. Retrieved from https://dspacecris.eurocris.org/handle/11366/393

Davis, F. D. (1989). Perceived usefulness, perceived ease of use and user acceptance of information technology. Quarterly, 13 (3), 319–340. https://www.jstor.org/stable/249008

de Castro, P. (2018). The role of Current Research Information Systems (CRIS) in supporting Open Science implementation: the case of Strathclyde. ITLib, (5), 21–30. https://doi.org/https://dx.doi.org/10.25610/itlib-2018-0003

de Castro, P., Shearer, K., & Summann, F. (2014). The Gradual Merging of Repository and CRIS Solutions to Meet Institutional Research Information Management Requirements. CRIS 2014, 13–15 May 2014, Rome, Italy. https://doi.org/doi:10.1016/j.procs.2014.06.007

Dijk, E., Hogenaar, A., & van Meel, M. (2006). NARCIS. Integrating CRIS, OAI and Web Crawling. CRIS2006: 8th International Conference on Current Research Information Systems (Bergen, May 10–13, 2006). Retrieved from https://dspacecris.eurocris.org/handle/11366/328

Engelman, A., Enkvist, C., & Pettersson, K. (2018). A FAIR archive based on the CERIF model. CRIS2018: 14th International Conference on Current Research Information Systems (Umeå, June 13–16, 2018). https://doi.org/10.1016/j.procs.2019.01.076

Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., … Hilf, E. (2004). The Access/Impact Problem and the Green and Gold Roads to Open Access. Serials Review, 30(4), 310–314. https://doi.org/doi: 10.1016/j.serrev.2004.09.013

Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). The fourth paradigm. Data-intensive scientific discovery. Redmond WA, Microsoft.

Ivanović, D., Ivanović, L., & Dimić-Surla, B. (2014). Multi-interoperable CRIS repository. CRIS 2014, 13–15 May 2014, Rome, Italy. https://doi.org/10.1016/j.procs.2014.06.014

Jeffery, K. (2012). CRIS in 2020. CRIS2012: 11th International Conference on Current Research Information Systems (Prague, June 6–9, 2012). Retrieved from http://dspacecris.eurocris.org/handle/11366/119

Joint, N. (2008). Current research information systems, open access repositories and libraries. Library Review, 57(8), 570–575. https://doi.org/doi:10.1108/00242530810899559

Kollmann, T. (1998). Akzeptanz innovativer Nutzungsgüter und -systeme: Konsequenzen für die Einführung von Telekommunikations- und Multimediasystemen, Gabler Verlag | Springer Fachmedien Wiesbaden GmbH, Wiesbaden. https://doi.org/10.1007/978-3-663-09235-3

Labastida i Juan, I. (2015). Managing the bridge between institutional repositories and CRIS: Universitat de Barcelona. EuroCRIS Strategic Meeting, Barcelona, 10th November 2015.

Lucke, B. (1995). Akzeptanz: Legitimität in der 'Abstimmungsgesellschaft', VS Verlag für Sozialwissenschaften | Springer Fachmedien Wiesbaden GmbH, Wiesbaden. https://doi.org/10.1007/978-3-663-09234-6

Lynch, C. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. Retrieved from http://www.arl.org/resources/pubs/br/br226/br226ir.shtml

MacGregor, G. (2019). Repository and CRIS interoperability issues within a 'connector lite' environment. *OR 2019 14th International Conference on Open Repositories*, Hamburg, June 10-13,

2019. https://pureportal.strath.ac.uk/en/publications/repository-and-cris-interoperability-issues-within-a-connector-li

Ribeiro, L., de Castro, P., & Mennielli, M. (2016). EUNIS – EUROCRIS joint survey on CRIS and IR. Final report. Paris, EUNIS. Retrieved from http://www.eunis.org/wp-content/uploads/2016/03/cris-report-ED.pdf

Rybinski, H., Skonieczny, L., Koperwas, J., Struk, W., Stepniak, J., & Kubrak, W. (2017). Integrating IR with CRIS – a novel researcher-centric approach. Program, 51(3), 298–321. https://doi.org/doi: 10.1108/prog-04-2017-0026

Rybinski, H., Kubrak, W., Skonieczny, L., Koperwas, J., & Struk, W. (2018). Omega-Psir – Institutional CRIS at Polish Universities. ITlib, (5), 36–44. Retrieved from https://itlib.cvtisr.sk/archiv/2018/5/omega-psir-institutional-cris-at-polish-universities.html?page_id=3532

Schöpfel, J., Azeroual, O., Saake, G. (2019). Implementation and user acceptance of research information systems. *Data Technologies and Applications*, Vol. ahead-of-print No. ahead-of-print. https://doi.org/10.1108/DTA-01-2019-0009

Siciliano, L., Schmidt, S., & Kinzler, M. (2014). BoRIS and BIA: CRIS and institutional repository integration at the Free University of Bozen-Bolzano. CRIS 2014, 13–15 May 2014, Rome, Italy, 33, 68–73. https://doi.org/10.1016/j.procs.2014.06.011

Suber, P. (2012). Open access. Cambridge MA, MIT Press. Retrieved from http://mitpress.mit.edu/books/open-access

Wenaas, L., Karlstrøm, N., & Vatnan, T. (2012). From a national CRIS along the road to green open access - And back again: Building infrastructure from CRIStin to institutional repositories in Norway. CRIS 2012. 11th International Conference on Current Research Information Systems, June 6–8, 2012, Prague, Czech Republic. https://dspacecris.eurocris.org/handle/11366/116

Venkatesh, V., Speier, C., & Morris, M.G. (2007) User Acceptance Enablers in Individual Decision Making About Technology: Toward an Integrated Model. *Decision Sciences*, 33(2), 297-316. https://doi.org/10.1111/j.1540-5915.2002.tb01646.x

Acknowledgments

The authors would like to thank Dragan Ivanovic (University of Novi Sad) and Martin Brändle (University of Zurich) for helpful comments and advice.