14th International Conference on Current Research Information Systems, CRIS2018

Practical CRIS Interoperability

Jan Dvořák¹,²,³,*, Tomáš Chudlarský¹,², Josef Špaček¹,†

¹Computing and Information Centre, Czech Technical University in Prague, Jugoslávských partyzáňů 3, CZ-16000 Prague 6, Czech Republic
²Institute of Information Studies and Librarianship, Charles University, Úklíže 8, CZ-15800 Prague 5, Czech Republic
³euroCRIS, Heyendaalseweg 141, 6525 AJ Nijmegen, The Netherlands

Abstract

Institutional Current Research Information Systems typically interoperate and interact with a rich variety of other information systems and services. We aim at providing a basic list and discussing the motivations, the technical aspects of the information interchange that is taking place, and potential benefits of applying CERIF. Our findings are illustrated on the example of the CRIS of the Czech Technical University in Prague, an in-house built CRIS with over two decades of history.

© 2019 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/3.0/)
Peer-review under responsibility of the scientific committee of the 14th International Conference on Current Research Information Systems, CRIS2018.

Keywords: Institutional Current Research Information Systems; information system interoperability; information interchange standardization; CERIF

1. Introduction

Higher education institutions have two main agendas: education and research. The research is documented in a Current Research Information System, or CRIS. Aside from that, every institution needs a finance information system, a human resources system and possibly other systems as well. Aspects of information about research are found in many of these systems, and in external information systems, too. A CRIS, in order to be effective, needs to cooperate, including interchanging information, with the other systems both inside and outside of the organisation. A CRIS can even be viewed as a “central relating information system”⁵.

The importance of interoperability of CRIS with other systems was recognized and studied in the last decade. Interconnection with literature repositories was studied²,⁸,¹⁰, aggregation to funder level¹, semantic questions were
discussed in the context of interoperability\textsuperscript{7,11}. Even minor integrations are found to increase the usability of CRIS\textsuperscript{9}. Integration and interoperability of CRIS was recognized as one of the core benefits\textsuperscript{8}. Integration of CRIS with surrounding systems decrease the administrative burden of the researchers, makes the data more valuable and consequently, supports the trust and transparency in the research domain (see also\textsuperscript{1}).

When interfacing with other systems, CRIS developers and managers have to deal with several types of heterogeneity: the heterogeneity in the underlying technology, in the data model, the semantic gaps, the different processes – these are the forms of heterogeneity of information coming from different sources. The goal is to reconcile the different views, hopefully resulting in a single, coherent picture.

In this paper the interfaces between the CRIS and the other systems will be discussed; the motivation, the possible technical realizations and an example one will be described. One will assess the potential benefits of having the information interchange happening in the framework of a CERIF profile.

2. The CTU CRIS

The Czech Technical University in Prague (CTU) is the largest technical university in the Czech Republic. It was established in 1707. Twenty thousand students are educated at eight faculties and six university institutes. The university has approximately three and half thousand full time equivalent staff.

The university has an in-house built CRIS of a long tradition. It started around 1995 i.e., at times when no products were being offered. Since then the institution hasn’t come across a system that would be worth the switch.

The CTU CRIS itself is composed of two applications V3S and EZOP (both working on a single database schema). The current technology is Java Enterprise Edition.

Metadata about outputs of research are being managed in the V3S application, which is the most recent addition to the system: it was put in operation on July 1st, 2016. Its predecessor started in the 1990-ies as an Oracle Forms application, which was later rewritten as a PHP application. The typology of research outputs maps onto the output types set recognized by the national research output database RIV, but is much more detailed. It thus accounts for discipline specificities, e.g. the various architectural design competition submissions that are necessary to faithfully describe the output of the Faculty of Architecture of the CTU. This component also tracks recognitions and service to the scientific community.

Information about projects and their funding is handled in the EZOP application. It was put in operation in 2015. While its predecessor system had limited functionality, project information in the institution has been centralised since the 1990-ies, too. This also includes contractual research projects.

The links between the research outputs and the research projects the outputs originated from are handled in the V3S system.

The CRIS, as well as most other systems, has consistently been built using an Oracle database. The use of a single database management system makes integrations with other internal systems at the data layer possible. These integrations are conveniently implemented through the database link mechanism and a set of database views on the side of the source system.

The scales of the main entities are the following:

- The organisation’s structure comprises 800 units.
- There are approximately 150,000 total persons with some sort of relationship to the CTU; of that, \(\sim22,000\text{–}25,000\) are active ones. (These numbers include all students.)
- Around 20,000 projects are tracked in the system.
- There are over 180,000 research outputs.

3. Institutional CRIS integrations with other information systems

We describe the different integrations of a CRIS. We discuss the following aspects:

(1) Motivations
(2) Information scope
(3) Technical aspects
(4) Potential benefits of applying CERIF
In (1), (2) and (4) we aim at general validity. The technical aspects in (3) are too varied and diverse so we present the particular case of the CTU CRIS as an illustration.

3.1. Article-level and citation-level bibliographic databases

The two most prominent services here are the Web of Science (WoS) by Clarivate Analytics and Scopus by Elsevier, with Crossref growing in importance. The Initiative for Open Citations – I4OC – aims at providing the links of the citation network.

Motivation: Pulling publication metadata from a bibliographic database minimizes manual (re)typing of the article metadata or avoids it altogether. It thus saves work and acquires verified article metadata and possibly also a finer categorization of publications. By tracking citations one gets the citation metrics that can be used – when put in the right context – for a bibliometric evaluation of the publications.

Information scope: Publication metadata, links between publications.

Technical aspects: The integration is most commonly done through the webservice API the database providers offer. Web of Science has a SOAP-style API, whereas both Scopus (and Crossref) offer a REST API, which is somewhat easier to use. For WoS and Scopus we do the following kinds of querying:

1. In a regularly scheduled nightly job:
   a. Search the database for recent (2 weeks) additions by the institution identifier (Organization Enhanced in WoS, Affiliation ID in Scopus) which consolidates all the various spellings of the institution’s name and address from the publication author affiliations.
   b. Check the continued presence and metadata of all publication identifiers from the given source that have been seen so far. This job walks through a portion of all the publications (dimensioned by the weekly request quota and/or a maximum request rate and the limitation of reasonable running time). Newer publications are checked more frequently than older ones. Currently it takes 4–8 days to check the publications from this and the previous years, and approximately 6–8 weeks to check the biggest tier of the oldest publications. It sometimes happens that an identifier of a publication changes in the external database. The record with the original identifier then becomes unavailable. Currently the bibliographic database vendors do not convey information on a transition of one record into another (although such a redirect instruction would in fact fit quite seamlessly in a REST API). The synchronization job detects this situation and invalidates the record in the CRIS database. This does not happen on the first unavailability though; we rather institute a quarantine period of 2 weeks, requiring at least 10 failed attempts in a row.

2. When a user enters a WoS or Scopus identifier that has not been seen so far, an ad-hoc background job is set up to contact the external database and retrieve the article information.

This way we keep a mirror of the publications from the external database in the CRIS. This is used to:

1. Offer researchers a starting point for creating their new records. The bibliographic information is copied, the user only has to pinpoint internal authors and their departments and to link the relevant projects and other funding.

2. Offer suggestions for linking an existing CRIS record with the external one (by filling in the relevant identifier: WoS code, Scopus code). The recommendation is based on matching DOIs and similarity searches in the titles.

3. When linked with the CRIS publication records, the external records are used to highlight differences in the metadata items. Where such differences are found, they are presented to the user who can embody them in the record or choose to ignore them (preferably recording the reason in a comment field for future reference).

4. When linked with the CRIS publication record, the external record pinpoints a position in the citation network that is tracked by the citation database. Citations are also mirrored from the external database in the CRIS. Researchers approve the individual citations and can correct the automated logic of recognizing self-citations.

We plan to integrate Crossref data and the data from the Initiative for Open Citation as well.
Potential role of CERIF: The bibliographic metadata is very well amenable to being represented in CERIF. CERIF would also bring the needed structural unification of such information. Notwithstanding, each database provider can follow their own document types and source types classifications.

3.2. Journal-level citation metrics sources

The Journal Citation Report by Clarivate Analytics and the Scimago Journal Rankings offer valuable sources of information about journals. Both the metrics (Journal Impact Factor, Scimago Journal Rank, Article Influence Score, Source Normalized Impact per Paper) and the discipline categorisations are linked to journals identified by ISSN and/or e-ISSN. Journal ranking of research publications is typically used in internal evaluations of organisation units and in academic promotion criteria. Internal reports combine bibliometric information at three different levels: journals, articles and citations. Where journal-level metrics are used, they are strictly considered within the discipline category of the journal and the year of publication.

Motivation: Finer categorization of articles, bibliometrics.
Information scope: Journal basic data, metrics, discipline categories.
Technical aspects: Ad-hoc imports (several times a year) of MS Excel files.
Potential benefits of using CERIF: Unification and possible automation of the ingest procedure, better tracking of changes in titles, publishers or ISSN's.

3.3. The institutional Open Access repository

Motivation: single entry of metadata, information on the Open Access aspects of publications also available in the CRIS
Information Scope: Publication and research dataset metadata, full texts and/or data files, rich context of the information assets (e.g. authors/creators, their detailed affiliations, research projects, research infrastructures).
Technical aspects: The interface needs an update for the refreshed CRIS.
Potential benefits of using CERIF: Transfer of the rich information context around a publication.

3.4. ORCID

The Open Researcher and Contributor IDentifier (ORCID) system is a global researcher profile registry that assigns public identifiers to researchers. This is unlike most other identifiers, which are not considered public: the institutional person identifiers for all internal systems of the university, the national person identifiers for communication with the national CRIS. ORCID identifiers are encouraged for their non-commercial status.

Motivation: By recording the ORCID iD of the researcher or doctoral student, the institution can trace the researcher’s activities from the time before they joined the institution, and possibly also after they leave it.

Information scope: ORCID offers two application programming interfaces (APIs): a public, read-only API and a members-only API that allows the researchers profile to be amended and modified. Both of these APIs consist of research information objects such as organisations (for education and employments), research projects, publications as well as other research outputs that map very well onto CERIF entities. Unfortunately, unique assignment of ORCID iDs to persons is not guaranteed.

Technical aspects: The CTU identity management system supports creating ORCID iDs or getting them through logging into https://orcid.org. The ORCID identifier is provided to other university systems along with user data. No further integration between the ORCID system and the CTU CRIS is done at the moment. Updating the ORCID record with publication metadata from the institutional CRIS is considered counter-productive. The opposite flow, of publication metadata from ORCID into the institutional CRIS might be useful, but bibliographic databases are more likely to contain more authoritative information.

Potential role of CERIF: The possibilities for a CERIF-based interface to ORCID by means of an adaptor were explored in. CERIF was found viable in principle, but currently lacking means of expressing the different visibility levels (public/members-only/private) at the level of individual fields.
3.5. Identity Management systems

With their wide user bases, CRISs at higher education institutions need to cooperate with the institutional user account policies and systems.

Motivation: Modern institutional information systems typically integrate their authentication function with an institutional login service. Also, the authorizations derive from the set of roles the user is playing in the organisation. Academic institutions typically use identity management systems (IdM) that centralize the management of roles across the whole institution.

Information Scope: The authentication side of the integration is handled by the institutional single sign-on service using established protocols and technologies. The authorizations are, however, evaluated inside the CRIS. The CRIS therefore needs a mirror of the relevant information from the identity management system:

1. The organisation structure of the institution;
2. The user identities (which include the names and titles, the institutional ID, the national ID and the ORCID iD of the person);
3. The detailed affiliation(s) of the person and the type of the relationship (employee or student);
4. The business roles of the person, such as head of department, dean, vice-dean for R&D, rector, etc.;
5. The technical roles of the person, such as the administrator of projects at a specific department or faculty.

Technical aspects: The CTU CRIS is accessible to the institution’s staff and doctoral students. The user’s permissions are checked in many contexts in the CRIS. The definition of user permissions is kept in a specific part of the CRIS database which is updated along with deployments of new releases of the CRIS. The evaluation whether user \( U \) has permission to do \( X \) with object \( Y \) (typically a project or a research output) is done entirely in a stored function in the database. This allows this logic to be called both by the business logic in the CRIS application and also in database queries the CRIS makes (e.g. the query to list the projects the user is permitted to see, which typically excludes contractual research projects the user is not involved in).

The specification of the user permissions is done in a MS Excel workbook of many worksheets. The permission typically depends on the combination of the business roles of the user in the organisation, on the role of the user with respect to the particular object, and on the type and the state of that object. Sometimes the permissions come through a longer chain of relationships. E.g. a user may be permitted to edit a link between a publication and a research project if he/she is the head of the organisation unit where the principal investigator of the project is affiliated.

Potential role of CERIF: CERIF currently does not include elements that would specifically handle user permissions. Its general-purpose structure of the links between the Person and OrganisationUnit entities, as well as the flexible semantic layer, certainly can convey the underlying data for authorization that was listed above.

3.6. Human Resources information systems

The HR systems in fact typically feed the Identity Management systems so in many cases a separate interface into a Human Resources system is not needed.

Motivation: The information about the staff contracts is important for a complex picture of research in an academic institution. The information is – directly or indirectly – used in many reports a CRIS produces.

Information scope: If there is need for a separate interface, it is typically a rather thin one: the type of the employment contract and the full-time equivalent.

Technical aspects: The CTU CRIS gets the information about the type of the employment contracts and the full-time equivalent through the identity management system interface, which eliminates the need for a direct connection of the CRIS to the HR system. For the purpose of the CRIS one also needs the history of the relationships: the CTU CRIS builds this history by periodic polling of the identity management system and recording the start and end dates of each relationship.

Potential role of CERIF: The CERIF link between a person and an organisation unit perfectly maps the information about an employment relationship, including its start date and also an end date (for past relationships).
3.7. Student management information system

Student management information systems contain information that is critical to higher education institutions. A part of it is relevant to the CRIS, too, as it touches research and its context in the institution.

Motivation: To have ready the information about theses from the perspectives of the student, of the supervisor, of a consultant or of a reviewer.

Information scope: The theses (doctoral, masters or bachelor), the doctoral studies with their different stages.

Technical aspects: The CTU CRIS currently sources the information on the doctoral studies (all of them: the in-progress ones, the completed ones, and the dropped ones as well) from the study management system called KOS. This is realized through a database link into the KOS database that provides specific interface consisting of several read-only views for this purpose.

The theses metadata are currently not included in any automated information flow into the CRIS. However, the theses are being pushed into the institutional repository. Where relevant, the students or the supervisors have the option to enter the thesis in the CRIS manually.

Potential role of CERIF: CERIF can very well represent the theses (as a specific subclass of publications) and their context. If a study process is viewed as a sort of project, it can also be mapped on CERIF. If such an interface is standardized, the systems at either side could be switched without affecting the other side of the interface. This would lead to a greater flexibility.

3.8. Finance information system

Motivation: To keep track of the allocated research funds and their progressive spending.

Information scope: Selected aggregates from the institutional accounting.

Technical aspects: In the case of the Czech Technical University in Prague the interlinking is done at the database level. The FIS information is represented in materialized views that are refreshed hourly.

Potential role of CERIF: The information is naturally structured in a way that does not follow the CERIF objects. It is by amending selected parts of the accounting data to the research information that the two systems can effectively cooperate.

3.9. The institution’s Document Management System

By virtue of the legal framework of the eIDAS EU regulation, electronic documents and signatures are equivalent to paper-based documents and hand-made signatures. However, such electronic documents need to be kept in specialized document management systems in order to retain their legal validity. Funding contract and other administrative documents related to research are natural candidates for storage in the electronic form.

Motivation: Have the funding contracts and other administrative documents easily accessible to users who are involved in the management of research in the institution, as well as to selected researchers (for the projects they are participating in). Have the documents stored in the electronic form while retaining their legal validity.

Information scope: The document itself and its metadata.

Technical aspects: The Czech Technical University in Prague uses a third-party product (called CUL) to handle the storage of administrative documents. This includes documents related to research projects, such as funding contracts and their amendments. Some of these documents can be subject to internal approval processes on the side of the university. These approval processes are handled by the CTU CRIS. They typically follow the organizational hierarchy from the project’s principal investigator’s department upwards.

Potential role of CERIF: The more general application domain of Document Management Systems makes them unlikely to implement a CERIF-based interface (even though some CERIF base objects can be roughly recognized in the DMS problem domain).
3.10. Internal Funding Agency

In some cases the university acts a funder: this agenda is either handled in the CRIS itself, or the relevant information needs to be communicated.

Motivation: To effectively and efficiently distribute the funds that are available, to manage and evaluate the resulting projects throughout their full life cycle.

Information scope: Information about projects and their context.

Technical aspects: The CTU CRIS interfaces with a specialized system (called SGS) to handle internal grants to support scientific activities by doctoral and master students. The funds are provided by the Ministry of Education, Youth and Sports of the Czech Republic under the budget line of “specific university research subsidy” (allocated to all public universities in the country).

The SGS application is a CRIS on its own, namely a funder CRIS. It handles the following activities:
- Students and/or their supervisors propose research projects
- Proposals are approved by the department heads and the deans
- The submission process with a deadline is handled
- The submitted proposals are evaluated
- The awarded projects are managed
- Committees perform interim and final evaluation of the projects
- Reports for the Ministry are generated

Project information from the SGS is transferred into the CTU CRIS. The research output metadata are collected in CTU CRIS and transferred into the SGS.

Potential role of CERIF: In the cases where an institution has a different system for handling its funder role, a CRIS-to-CRIS communication takes place. The implementation of CERIF on this interface is a natural choice.

Note: The roles of a research funder and of a research performing institution are fundamentally different. Even in situations where some research funding distribution is delegated to a research performing institution, it makes sense for this institution to keep two different CRIS. However, information interchange between them is essential. Both agendas being handled in a single CRIS is another valid choice. The particular CTU setup is a result of a long and rich history.

3.11. The institution’s technology transfer office

Research performing institutions have to promote commercial exploitation of their results. This is typically done through technology transfer offices that will have information on license contracts for the institution’s applied R&D outputs and on spin-off companies.

Motivation: To effectively support the work of the technology transfer office, to collect and disseminate (while respecting IP rights, contractual obligations and interests of the stakeholders) information about actual or potential technology transfer.

Information scope: License contracts, spin-off companies, researchers pursuing their idea into production, other forms of technology transfer.

Technical aspects: The CTU CRIS currently keeps track of patent licensing contracts, including the amounts. This information has limited visibility.

Potential role of CERIF: If the information is kept or needed in a different system, CERIF is the natural communication format.

3.12. Funder CRISs

The interface between the institutional CRIS and the CRISs of funding agencies has a great potential for interchange of machine-processable information. The current level of integration on this front seems, however, very low.

Motivation: Sharing the information of a research project proposal between the funder and the involved institutions in the proposal preparation would significantly reduce the workload of the researchers and allow for
much more efficient internal approval processes on the side of the applicant organisation(s). Communicating the research project proposal evaluation outcome from the funder to the institutions in a structured way would greatly simplify the project setup processes. Structured, semi-automated interim and final reporting on a funded project (i.e., the activities undertaken, the costs, and the resulting outputs) has the potential to greatly reduce the administrative workload on researchers.

Information scope: Project information including its context (organisations, researchers, funding) communicated both ways; reports (including research output metadata) communicated from the institutions to the funders.

Technical aspects: This is currently a blue-sky integration. We are not aware of any established, automated information interchange between institutional and funder CRISs anywhere in the world. The current way of communicating is through electronic forms and letters. If these could be augmented with the same information represented in a machine-processable way, they could be used for both person-to-person and CRIS-to-CRIS communication. Technically, this could be realized as PDF documents with CERIF-XML attachments (embedded in the PDF file); authenticity could be established by means of electronic signatures or electronic seals complemented with electronic timestamps.

Potential role of CERIF: CERIF is the natural communication standard for such information interchange.

4. Reporting

The core of the value an institutional CRIS brings to its institution is in the reports it can generate. This includes reports on research inputs, on the research process, and on the outputs of research. The reports can be internal only, shared within a group of institutions, generated for a particular external party (a funder, statistics office or other supervision body, a research information aggregator such as a national CRIS or OpenAIRE), or public. Reports can list different aggregation levels, from the whole institution down to individual researchers.

Internal reports include bibliometric or other reports for evaluations (both internal and external ones, at different granularity levels with different indicators and perhaps with discipline specificities), supportive information for academic promotions, insights for the strategic management of research at the institution.

Some of the reports allow research activities to be monitored in the broader context of the institution. Some institutions even build an institutional data warehouse that consolidates information from all parts of the institution: research, education, administration.

The CTU CRIS generates a diverse scale of reports. This includes:

- Internal evaluation reports for departments of some faculties;
- Reports that track the fulfilment of criteria for academic staff promotions in the ranks of associate or full professors;
- Exports of research output metadata for the national RIV database;
- Reports on R&D for the national statistics office.

Where a report does not exist yet, one can be generated ad-hoc by directly querying the database. An internal data warehouse is in an early phase of implementation.

5. Conclusions

There are many diverse systems a CRIS needs or can integrate with. The interfaces can have various styles: webservice APIs, database integrations (internally), data file interchanges, in some cases even web harvesting. The external information, once collected, needs to be disambiguated. This is relatively easy in the parts where identifiers are available. It gets much more complex where fuzzy string matching or other similarity recognition techniques need to be employed.

An in-house built CRIS has the advantage of a greater flexibility in setting up interfaces with the neighboring systems. It can be more tailored to the needs of the institution, accounting for the organisational or national specificities. This flexibility is limited for all the other options of acquiring a CRIS: contracted bespoke-built systems, CRIS products, or a Software-as-a-Service CRIS offerings. We therefore argue that an in-house built CRIS is better integrated into the information ecosystem of the institution. External interfaces of a CRIS, however,
represent a maintenance burden. Their standardization would play a big role in keeping the maintenance costs bearable.

We assess the potential of CERIF as the basis of such standardization. CERIF is a natural standard for interchange of research information. This typically includes CRISs, even if they are of different types: a bibliographic database, a researcher profile system, or a project proposal submission system of a funder. Where the core of the information interchange is different (such as in a finance information system or a document management system) it is much more productive to follow established standard formats and protocols for the problem domain.

CRIS integrations are a prerequisite for usability of the CRIS. In effect this saves researchers’ time and provides management with a complex view of research at the institution. We see this as an essential feature of a successful research-oriented university.

References