

Preparing for CRIS: Challenges and Opportunities for Systems Integration at Finnish Universities

Miika Nurminen, University of Jyväskylä, Finland, minurmin@jyu.fi

Abstract

This paper presents issues identified on the attempt to integrate administrative research information to institutional repository (IR) and other systems, considering local requirements and national publication reporting process managed by the Ministry of Education and Culture. The observations are based on preparing for procurement of a Current Research Information System (CRIS) at the University of Jyväskylä. The CRIS will be used by various stakeholders in different organizational units, having conflicting requirements and different notions on system usage (e.g. publication reporting, project management, researcher CVs). Determining the optimal data flow for handling publications, organizations, or financial information in different systems needs architectural consideration.

While it would be preferable to integrate CRIS and IR to a single system, there are issues with user interface, recording conventions, and selecting the data to be included that may make the task impractical. However, both systems will benefit from synchronization of selected datasets and separation of responsibilities. More generally, Finnish universities would benefit on services that assist on federated handling of publication data (e.g. publication forum class calculation, determining collaborative publications, transformations for data import/export). The overall publication reporting process is problematic in its current form and may need revisions at the national level.

Audience: Research administration, IT administration, librarians, repository managers.

1 Background

Finnish universities are undergoing a fundamental change related to their legal position and governmental funding. Since 2011, the universities have had to provide full publication metadata to the Finnish Ministry of Education and Culture. The focus in publication reporting has shifted from summary statistics to emphasizing high-quality publication channels as determined by ranking in the Publication Forum (Auranen & Pölönen, 2012). Even though there were plans to construct a national publication database to manage the metadata in a centralized way (Ilva, 2011), it never came to pass, necessitating considerable development efforts in individual organizations. Currently, many universities in Finland are (or already have) procuring a commercial CRIS system. As reporting requirements become more complex, the responsibility to record the publication data is shifting from researchers to the university libraries.

Whereas in-house development of research information systems is becoming rare, following the trend in UK (Russell, 2012) and other European countries, institutional repositories based on open source and local customizations have been well-established to support self-archiving. As open access mandates become more commonplace, it seems likely that the library staff will take more active role in the self-archiving process (as part of recording the publication metadata in CRIS) as well. This paper addresses the issues presented by integrating administrative research information to institutional repository and other systems, based on observations within the TUTKA working group at the University of Jyväskylä (Miettinen et al., 2013), along with lessons learned while maintaining the legacy research information system TUTKA¹ and institutional repository JYX².

¹<http://tutka.jyu.fi/>

²<https://jyx.jyu.fi/>

2 Considerations for System Development

The working group was commissioned to investigate implementation options for the future research information system. Two major commercial products were evaluated: Elsevier's [Pure \(2011\)](#) and Thomson Reuters' [Converis \(2013\)](#). Open source alternative DSpace-CRIS³ (=with the prospect of integrating it to JYX) was also considered, as well as the option to modernize the legacy system developed in-house.

2.1 Organizational Context

TUTKA group identified four types of entities that are regarded to be essential and validated *research* information system components: *publications*, *travel information*, *pre- and post-awards management* (projects), and important scientific *prizes*. While the current system contains other data of interest as well (e.g. presentations, thesis supervision, other scientific activity), it was not regarded as critical in the research management sense (obviously, the other data does have relevance as part of a CV-system or *study*-related information systems) - the point is to make a clear distinction with the responsibilities of the different systems and minimize the amount of overlapping data entry.

There are certain features that are of special importance when procuring the new system. These include the ability to link the research-related entities and other activities to each other (CERIF⁴-like data model), and to define clear, centralized workflows for recording research-related information such that the amount of administrative work is minimized. For publications, the library takes care of recording publication data, utilizing external bibliographic databases. For projects, the researcher must submit a description of the project for department head even before applying funding. All the relevant project documents (e.g. funding applications and decisions) will be archived to the CRIS, usable by all the stakeholders.

2.2 Accommodating National Publication Reporting Requirements

Until Spring 2014, the task of recording publication data was mainly left for researchers. There was a *publication responsible* person in every department, ensuring that the data gets marked in accordance to the requirements posed by the Ministry. Ever since the reporting of full publication metadata started, this task has become increasingly difficult as the number of classification rules and obligatory fields keeps increasing. The most serious problem is that the concept of "publication" in the sense of national requirements is becoming less obvious to the researchers – and sometimes even to librarians as well:

- Edited or translated books can be attributed to the editor only if the book has "extensive introduction containing editor's own scientific contribution".
- Determining whether an edited book is a "report" or "scientific book" radically affects the convention used to mark the chapters: if the publication is classified as a report and all authors are from the same organization, the chapters are ignored and the whole report is treated as a single publication.
- Marking conference articles is often confusing. Depending on the type of proceedings, the papers may end up as *journal* articles, or even plain book chapters. If the articles are published only electronically and do not resemble enough a traditional, printed publication, they might be excluded.
- The Ministry collects increasingly detailed information about national collaborative publications (e.g. written with other Finnish research institutions). Determining these efficiently is problematic since the detail of affiliation data varies and some articles may have hundreds of authors.

In the technical sense, the publication reporting process as a whole is problematic because of its distributed, one-way nature: for collaborative publications, multiple HEIs (38 universities and polytechnics) end up reporting the same data that is checked, combined, and possibly sent back to original institutions in case of conflicts (e.g. ambiguous publication type). After the data is reviewed and assigned publication forum levels, it is imported to national publication portal Juuli⁵ – at this stage, erroneous data can no longer be corrected. The publication forum has issues as well, since the exact way to calculate the level for varying publication types and spellings can

³<http://cineca.github.io/dspace-cris/>

⁴<http://www.eurocris.org>

⁵<http://www.juuli.fi/>

be ambiguous. Even though the basic data is available in CSV lists, the final levels for publications are known only after the data is reported, since new rankings are added based on the data. In the future, more dynamic update process (including state research institutes and university hospitals) may take place (Haapamäki, 2013), but with little specifications available, this cannot be yet properly accounted for in the CRIS requirements.

Because of the "impedance mismatch" between national reporting requirements and researchers' own concept of publication lists or CV data required by funders (and even university's internal research assessment), the ability to mark *all* publications – even if not counted in official reporting – is an important requirement for the CRIS. Even adopting the CERIF model would not solve the issues (although it would help with sharing the publication data between different systems prior to actual reporting). The Finnish XDW reference model⁶ has been developed separately from CERIF with initial focus on educational data. While having many similar concepts, transformation rules need to be generated (Kaitera, 2013). Besides, the primary authority for publication data is the Data Collection Manual⁷ defined by the Ministry – with final clarifications given sometimes very close to the time the publication data is supposed to be ready for reporting.

2.3 CRIS-IR or IR-CRIS?

Clements (2013) argues that the question of "choosing" CRIS over IR is irrelevant: "stop thinking about systems and start thinking about services". While providing services is undeniably essential and we certainly agree about the principles of good information management outlined in the article, in order to *implement* the services effectively, it is critical to understand the relative strengths of the respective systems. For example, CERIF-compliant systems have a formal schema definition and better validation options to more loosely-defined insitutional repositories (Asserson & Jeffery, 2010). In general, this makes a conventional IR unsuitable as a basis for CRIS – despite the obvious overlap in metadata and basic functionality.

From the architectural point of view it would be preferable to integrate CRIS and IR to a single system – considering both system maintenance, and recording effort. The fact that while procuring a new system, this option allow us to actually *eliminate* a system by consolidation (and consequently, providing one primary "window" to university's research activities instead of multiple different interfaces showing partially the same data). Commercial CRIS vendors have extended their systems with built-in IR functionality (Pure, 2011; Converis, 2013), but in our case, replacing the IR with a commercial CRIS is not an option, since JYX contains a multitude of material that is already used in other services, or would not be suitable as CRIS entities. This includes digitized photographs, maps⁸, music data, and other archival material. The option of extending institutional repository to CRIS shows more potential since it facilitates managing the full publication recording workflow in the library (instead of only parallel publishing or library cataloging). The prominent example is DSpace-CRIS – used impressively in the HKU Scholar Hub (Palmer, 2010). It provides a customizable data model, CRIS entities as authority for item metadata, and a framework for retrieving publication data from bibliographic databases (Bollini & Mornati, 2013). Recently, DSpace-CRIS has even been extended with CERIF export (Mornati & Bollini, 2013). While building a new research information system from scratch is not a feasible option with current developer resources, extending an open source solution deserves more consideration.

Unfortunately, according to preliminary tests, it seems that DSpace-CRIS would need extensive tailoring in order to be useful in our context. Even though the publication module could be adequately extended (there may be issues with representing author affiliations in publications efficiently), the submission workflow is problematic, since it is centered around files and less about concepts and relations *about* the publication (XML workflow is not compatible with JSPUI used by DSpace-CRIS). Standard forms provided by DSpace are too limited concerning validation and flexibility to handle different publication types in a user-friendly way (using CRIS entities for authority control helps to certain extent). For other CRIS entities, UI issues are alleviated with JDynA framework, but the problem with workflow becomes even worse. The complex process of pre- and post-wards management would require a new kind of workflow. The extent of customization (and synchronizing the changes with the main development trunk) makes DSpace-CRIS less attractive option for data entry. However, it could be useful as a CV Portal, provided that the data is edited in other systems and imported to DSpace.

The investigation showed that a commercial CRIS would most likely be sufficient for our needs, provided that the requirements are set rigorously. While both major commercial systems (and possibly DSpace-CRIS-

⁶<http://tietomalli.csc.fi/>

⁷<https://confluence.csc.fi/download/attachments/21072701/Publication+data+collection>manual+universities+2012.pdf>

⁸<http://www.vanhakartta.fi/>

based IRIS⁹) offer a good basic set of functionality (e.g. publication data import, flexible workflows and comprehensive reporting), there are many concerns, such as the customizability of the system. It is essential that the data model (and consequently, validation of input forms) can be modified without the intervention of the vendor, since new reporting and validation requirements may arise in a short notice. In addition, the flexibility of interfaces regarding both data input (i.e. MARCXML import for publications in JYKDOK library system; financial and travel import from SAP) and connectivity to external web services are considered important.

3 Conclusion

Integrating CRIS and IR to a single system would offer obvious advantages. However, there are issues with user interface, recording conventions, and maintenance cost that may make the task impractical, especially considering the data entry and workflows required for pre- and post-awards management. The fact that similar data exists in multiple systems is not a problem provided that the responsibilities of the systems are clearly defined and there is just one entry point for certain types of data. This opens opportunities to streamline also related processes: for example, student theses management spans multiple organizational sectors, but the natural place for master data would be the study data system (e.g. Korppi¹⁰). Thesis data could be imported to CRIS (supplemented with other information such as the related research project) to support performance assessment, and finally to JYX and even JYKDOK as metadata related to thesis files. The question of integration should not be limited to IR or CRIS: generating reports and collecting the data from different sources is shifting from individual systems to a centralized data warehouse. The increased significance of research datasets and related source code pushes the requirements for storage and scalability yet further.

Even without a national publication registry (cf. CRIS in Norway; Sidselrud & Lingjærde, 2012), Finnish universities would benefit on services that assist on federated handling of publication data. Some development projects have already taken steps towards this direction (e.g. plans for centralized publication forum calculation service; national service bus; extensions to national article registry ARTO¹¹), but yet more collaborative effort that allow entering cross-institutional data only once prior to national reporting is needed. This does not exclude using university-level CRIS as well, as realized by DiVA (Andersson et al., 2013) portal in Sweden.

References

- Andersson, S., Klosa, U., Sundin, M., & Svensson, A. (2013). *DiVA: A well rooted and growing platform*. (LIBER 42nd Annual Conference)
- Asserson, A., & Jeffery, K. (2010). CRIS and institutional repositories. *Data Science Journal*, 9.
- Auranen, O., & Pölonen, I. (2012). *Classification of scientific publication channels – final report of the Publication forum project (2010-2012)*. Federation of Finnish Learned Societies.
- Bollini, A., & Mornati, S. (2013). *Integrate external bibliographic services in DSpace submission process to make self-deposit easy and improve metadata quality and presence of fulltext*. (Open Repositories 2013)
- Clements, A. (2013). Research information meets research data management. *Libr. Conn. Newsletter*, 11(1).
- Converis 5.1 fact sheet. (2013). Avedas.
- Haapamäki, J. (2013). *OKM:n julkaisutietojenkeruuprosessi ja sen kehittäminen* [Publication reporting process and its development by the Ministry]. (TUTKI-Publication data collection seminar, 18 Sep. 2013)
- Ilva, J. (2011). Building research assessment tools on a national level. *The National Library of Finland Bulletin*.
- Kaitera, A. (2013). *National CRIS development in Finland*. (euroCRIS Membership Meeting, 14-15 Nov. 2013)
- Miettinen, K., Auer, A., Kokko, J., Korppi-Tommola, S.-L., Liimatainen, S., Nurminen, M., ... Tuominen, K. (2013). *TUTKA-ryhmän loppuraportti* [TUTKA working group final report]. University of Jyväskylä.
- Mornati, S., & Bollini, A. (2013). *DSpace-CRIS: an open source solution*. (euroCRIS Membership Meeting, 14-15 Nov. 2013)
- Palmer, D. (2010). The HKU Scholars Hub; unlocking collective intelligence. *Chinese J. of Library and Inf. Science*, 4(1).
- Pure whitepaper 4.11.0. (2011). Atira.
- Russell, R. (2012). *Adoption of CERIF in higher education institutions in the UK : A landscape study*. UKOLN.
- Sidselrud, A., & Lingjærde, G. C. (2012). The practical implementation of the CRIS system CRIS in and the goals/challenges of bringing 150 institutions into production within a year. In *Proceedings of the 11th International Conference on Current Research Information Systems*.

⁹<http://www.cineca.it/en/content/iris-institutional-research-information-system>

¹⁰<https://korppi.jyu.fi/>

¹¹<https://arto.linneanet.fi/>