TOWARDS A CRIS-INTEGRATED SOLUTION FOR UNIVERSITY PRESS WORKFLOWS

Bastian Melsheimer[1], Annika Jahn[1], Markus Putnings[2], Valianos Stelica[2], Marcus Walther[3*]

[1] Präsidialstab / Friedrich-Alexander Universität Erlangen-Nürnberg
[2] Universitätsbibliothek / Friedrich-Alexander Universität Erlangen-Nürnberg
[3] Competence Unit for Research Data and Information / Friedrich-Alexander Universität Erlangen-Nürnberg
[*] Corresponding author: marcus.walther@fau.de; Schlossplatz 4, 91054 Erlangen

1 Introduction and motivation

At the Friedrich-Alexander Universität Erlangen-Nürnberg (FAU) the Converis software [1] is used as research information system (FAU CRIS, https://cris.fau.de/). The basic building blocks of FAU CRIS are personnel, organizations and functions associated with a business card. In this context, every FAU employee has at least one business card. Attached to the business cards is the research output, such as publications or research projects. In order to enable the continued use of the entered data and to simplify some process steps for the users and the administration, various CRIS-integrated workflows will be established in the system.

One example of implementation in the FAU CRIS was an application for university-internal promotions and funding for young scientists. The original application was paper-based. By using the research information system, the process steps are digitalised and thus simplified. This leads to a data-oriented procedure and increases efficiency.

Similar workflow improvements are currently planned in collaboration with the university library: previously, the full-text open access repository OPUS FAU and FAU CRIS were completely separate. In the past, the meta data for first and secondary publications in the repository had to be entered redundantly in FAU CRIS. By linking FAU CRIS and the repository, this process is improved and the green route of Open Access will be strengthened by integrating Sherpa/Romeo.

2 Initial situation

2.1 Library

The University Library has been using a hosted an OPUS 4 instance called OPUS FAU as an open access full text repository for the last decade. The focus was on publications, specifically first publications (e.g., dissertations and books from FAU University Press as well as FAU publication series which use OPUS FAU as their original publication platform) and secondary publications of articles in the green route of Open Access. In addition to secondary publications by authors themselves, publisher articles have also been fed in automatically via the DeepGreen[1] data hub since around 2016.

OPUS 4 itself is an open source software under the GNU General Public License, whose further development is largely in the hands of the German library network Kooperativer Bibliotheksverbund Berlin-Brandenburg (KOBV). In addition to providing the open source software, the KOBV headquarters also hosts OPUS 4 for over 46 institutions in Germany, making it one of the largest providers of repository hosting services in German-speaking countries [2]. The biggest advantage is the robust operation, the coverage of all basic needs (OAI-PMH and SWORD interface, support of a DOI assignment, etc.) and the cost-effective hosting.

The disadvantage of a hosted instance is the low flexibility, be it in terms of corporate design customization or in terms of functionalities and interfaces that go beyond the basic needs. In addition, a future-proof and adaptable extensibility of the metadata schema to other document types such as Open Educational Resources, research data, software, etc. would be desirable.

[1]DeepGreen (*https://info.oa-deepgreen.de/*) is a German delivery service for OA articles based on the software of the UK Jisc Publications Event Router (*https://www.jisc.ac.uk/publications-router*). In contrast to the latter, however, it always and exclusively delivers open access full texts and from several license contexts, i.e. in addition to pure gold open access articles, it also delivers articles from national licenses, alliance licenses, and licenses of German subject information services.

Taking into consideration the complex and fast changing environment in the research and publication field, we decided to have more flexibility and control on our open-access repository and started to look for a system that can be easily customized and developed. In the context of FAU's new Open Science Policy, which also places emphasis on open source, we wanted an open source repository that would allow local management of our digital content in any format (PDF, Word, JPEG, TIFF, etc.), that would be well supported by the community, and that would be based on a common technology. Furthermore, because we have many systems in our university that need to exchange information with the repository (e.g. CRIS, IdentityManagementSystem (IdMS), ORCID), a full REST API software was a high priority. This exchange of information will be filled with life as described in this paper. With this in mind, we opted for a DSpace based repository, which uses Angular for the frontend and Java, Solr and Oracle/PostgreSQL for the backend. DSpace has regular releases, several preconfigured features and is used in most of the universities[2].

2.2 FAU CRIS

FAU initially started work on its research information system CRIS in 2011 based on Converis 4 [3]. After the initial process specification, identification of possible source systems and an upgrade to version 5, it was introduced gradually from 2014 to 2016. CRIS includes metadata on publications, research projects, various other research output and supporting information on research fields, infrastructure, funding, etc. In 2019, regular automated imports from both Web of Science and Scopus were implemented for FAU-affiliated papers [4]. Depending on the discipline, the automated processing takes care of between 20% (Social Sciences and Humanities) and over 90% (Natural Sciences and Medicine) of the total peer-reviewed publication output.

One of the most important incentives to using the CRIS system is the WordPress plugin, which allows research information to be displayed on the chairs' and departments' websites throughout the university. The plugin was developed together with the Regional Computing Centre Erlangen (RRZE) and is updated regularly with customer's ideas and features [5].

Despite the long development and usage, FAU's library IT systems and workflows were not directly linked to FAU's CRIS implementation, which will be changed with the new workflows.

2.3 Data integration in the IT department

FAU Erlangen-Nürnberg IT services are managed by the Regional Computing Centre Erlangen (RRZE). This facility provides IT services not exclusively to FAU, but also to other universities and research facilities in Northern Bavaria.

A major challenge is data exchange between different software systems. Often there are neither suitable APIs nor sophisticated import and export interfaces. Usually the users of different university systems work only in their specific domain, while the information they administer is often needed elsewhere. Therefore, the process of data exchange is not only a technical task, but also an organisational challenge.

So, RRZE's "data integration platform" (DIP) is not just an IT system [6]. It includes requirements definitions and management for data flows. Each data flow requires three roles:

1. Customer
   - in authority of target system & with it's domain knowledge
   - needs source system's data
2. Source System Manager
   - domain knowledge on source system
   - offers data "as it is"
3. Data Flow Admin
   - technical / administrative competence in data flows
   - competence on aggregated data, APIs, database access, data transformation
   - in the IT department

[2]OpenDOAR (https://v2.sherpa.ac.uk/view/repository_visualisations/1.html) is the quality-assured, global Directory of Open Access Repositories. Each repository record within OpenDOAR has been carefully reviewed and processed by a member of our editorial team which enables us to offer a trusted service for the community.

DIP currently uses 28 source systems (including our CRIS system) and feeds 36 target systems (again including our CRIS). Therefore, this well-established platform at FAU was the logical choice for powering and linking the library's DSpace installation.

The typical challenges of IT projects like finding common language and definitions or frequent updates on requirements and workflows are addressed by a change and divergence aware software and user interface design.

The implementation of a new data flow consist of three steps:

0. preparation:
    * prioritisation and resource allocation
    * initial requirements definition
    * general setting
    * identification of source systems and it's data
    * get permission for source data access
1. prototype:
    * catalogue of requirements
    * interface definition (versioned, independently from implementation)
    * implementation of pilot data flow on QA system
    * testing and approval by Customer
2. productive
    * service and maintenance operation
    * move from QA to production
    * monitoring by system's operation

Technically DIP is an abstraction layer between different systems. Based on Grails infrastructure using MongoDB as Data Lake, it implements a domain specific language with custom extensions in Groovy [7, 8]. Data updates from source systems can be triggered per on-change-event or time based for whole datasets.

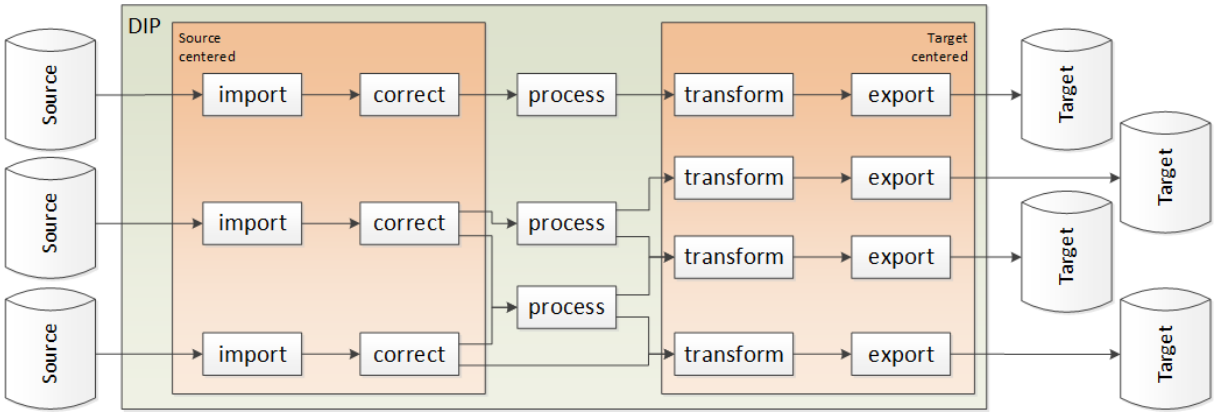

Fig 1: Data Flow Architecture.

Technically, each data flow contains three parts (see Fig 1). First is *source centred*: *Import* and *corrections*, like changing strings to numbers, fix known issues in source data. *Process* optimizes data handling and builds data domains inside DIP. Last is *target centred*, *transformations* prepare data for each target system (e.g. turn numbers back to strings for one target, but not for the others). Finally, *export* connectors handle the output.

Data integration demands a high level of data quality. DIP includes *data validation* logic, which informs Source System's Managers in case of validation errors. This is very helpful if the source system cannot identify such errors internally.

3 Creating a Dspace-CRIS connector

3.1 Previous related work regarding workflows

FAU CRIS already includes workflows incorporating different internal organisations. For example in 2018 the promotion for outstanding scientists to become head of a junior research group (prior being state doctorate, German "Habilitation") was implemented. The process starts with the research information already present in CRIS. After formal review from the university's administration the application is passed to the graduate centre and the responsible dean for the final decision. Except from filing the new status into the personnel file the whole workflow is maintained in FAU CRIS.

The Emerging Talents Initiative, as a second example, is a university-internal funding support for postdocs to write and subsequently submit a proposal financed by third-party funds. Through the implementation in FAU CRIS, some process steps could be optimised and the entire procedure, such as the application, review and also the subsequent final report can now simply be entered in the CRIS. A big advantage is that a lot of research information can be transferred easily and the scientist thus saves time. In addition, the scientists are immediately informed of mistakes and are able to correct them in a timely manner.

Accordingly, FAU has both theoretical-procedural experience in linking various organizational subunits and metadata/technical possibilities for doing so by means of DIP.

3.2 Work in progress

Previous experience with the aforementioned workflows and its digitization and optimization as well as the connection of existing data sources immediately showed that a direct connector, as offered by Converis, does not provide sufficient flexibility. This led to the decision to involve DIP and its services as "intermediaries" for the planned DSpace connector.

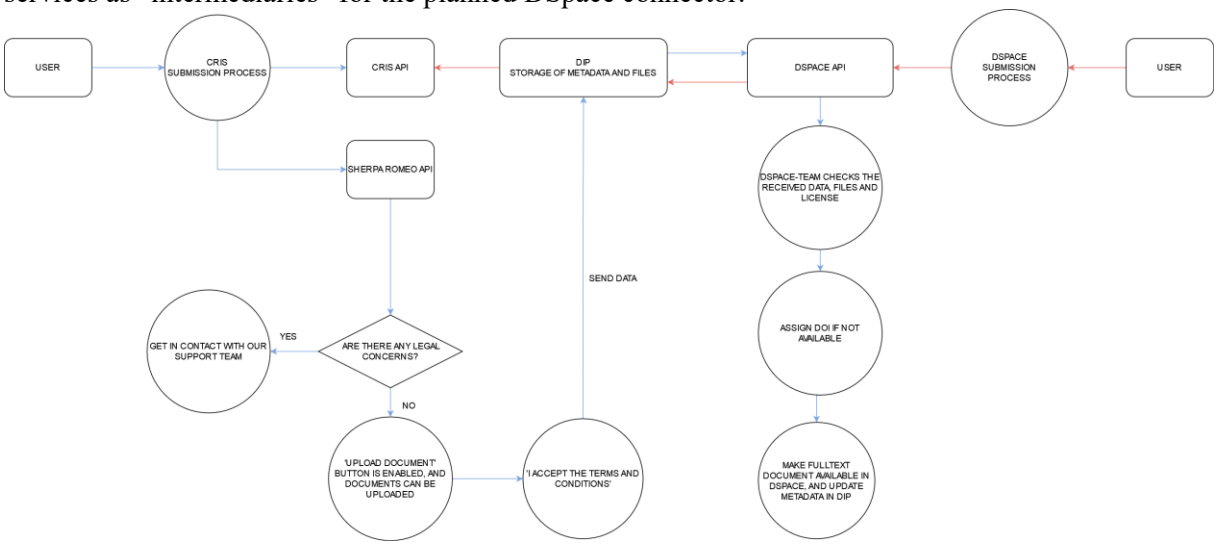

Fig 2: Workflow diagram.

A user can start the submission process in DSpace or CRIS and sync its upload with the other system (see Fig. 2).

Once the user fills in the submission form in CRIS (blue arrows in Fig. 2) and has the legal rights to publish a fulltext in DSpace (verified via Sherpa Romeo and with appropriate warnings), he can accept the terms of publishing in DSpace and then proceed with the upload. The files and the related metadata are transferred to DIP (Data Integration Platform), where further dissemination will take place, in order to map the metadata to DSpace and to provide additional information about the document. Furthermore, DIP will send the files and the related metadata to DSpace (via the REST API), and these will be available in the task pool, where the library's editors will check them one more time. If the documents have no DOI, DSpace will provide one using Datacite. The new DOI and other edited metadata are updated daily in DIP by a DSpace cronjob.

If the user uploads a document in DSpace (red arrows in Fig. 2) and wants to make its metadata

available in CRIS, the process is similar, but the DIP will receive only the metadata. It is also important to notice, that the users require a FAU IdMS account, which both systems use for the authentification process.

Implementation of the workflows shown in the diagram above is planned for 2022/23 and will be realized by a joint task force of library, IT and CRIS personnel. The process will start with published journal articles. As future options, books and book chapters will be supported, too, and an OER workflow for teaching material (and other systems, such as www.studon.fau.de/oer) would also be possible.

4 Conclusion

The main advantage of our setup is the decoupled data exchange between two standard software systems. There are no changes necessary in either Converis or DSpace. Workflows and software upgrades on each side can be done independently since the two systems are only loosely bound using FAU internal or even global PIDs. By adding the DIP as data exchange layer we gain more flexibility and reduced dependency on software development by the community and/or the manufacturer. We aim to broaden Open Access to research outcome through our updated user and data centred measures fitting ideally into FAU's new Open Science Policy [11].

5 References

[1] Maier, Jan C., Höllrigl, Thorsten and Weiss, Rudolf (2012): "CONVERIS 5: the Next Generation Current Research Information System."

[2] Rusch, Beate; Peters-Kottig, Wolfgang; Boltze, Julia; Brandtner, Andreas; Degkwitz, Andreas; Kirsch, Simona et al. (2021): KOBV Jahresbericht 2019-2020. doi: 10.12752/8247.

[3] Melsheimer, B., & Walther, M. (2016). Introducing CRIS at FAU. Procedia Computer Science, 106, 239-244. https://dx.doi.org/10.1016/j.procs.2017.03.021.

[4] Walther, M., & Melsheimer, B. (2019). Automated author affiliation processing using Scopus data. Procedia Computer Science, 146, 53-59. https://dx.doi.org/10.1016/j.procs.2019.01.079.

[5] https://www.wordpress.rrze.fau.de/plugins/fau-und-rrze-plugins/fau-cris/

[6] https://www.rrze.fau.de/medien-entwicklung/daten-systemintegration/datenintegrationsplattform-dip/.

[7] Abdul-Jawad, Bashar (December 2, 2008), Groovy and Grails Recipes (1st ed.), Apress, p. 424, ISBN 978-1-4302-1600-1.

[8] Chodorow, Kristina; Dirolf, Michael (September 23, 2010), MongoDB: The Definitive Guide (1st ed.), O'Reilly Media, p. 216, ISBN 978-1-4493-8156-1.

[9] Executive Board of Friedrich-Alexander-Universität Erlangen-Nürnberg. (2021). Guidelines for implementing open science at Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). Open Science Policy. Zenodo. https://doi.org/10.5281/zenodo.5602560 .