JOIN2 A PUBLICATION DATABASE AND REPOSITORY BASED ON INVENIO

L. Baracchi[1,a], A. Wagner[2,b] JOIN[2] Collaboration

[1] *Deutsches Zentrum für Neurodegenerative Erkrankungen, DZNE, Library*

[2] *Deutsches Elektronen-Synchrotron, DESY, Library*

E-mail: [a] laura.baracchi@dzne.de, [b] alexander.wagner@desy.de

JOIN² is a shared repository infrastructure that brings together eight research institutes for the development of a full-fledged scholarly publication database and repository based on the Invenio v1.1 open source framework for large-scale digital repositories. Seven JOIN² instances are already successfully deployed and two more institutes have joined seamlessly during the last year, resulting in the overall consolidation of the system and its functionalities. JOIN² provides a general solution built around a well-defined publication workflow which represents the cornerstone of the JOIN² paradigm. Always preferring simplicity to complexity and implementing a convergent, inclusive solution, the JOIN² members have consolidated their successful development workflow and collaboration. We highlight how JOIN² is able to address the needs of a heterogeneous group of research centres.

Keywords: JOIN², publication database, institutional repository, library system, publishing, services, reporting

Laura Baracchi, Alexander Wagner

1. Introduction

With the main goal of establishing a centralized single source for publication reporting, already in 2000 *Forschungszentrum Jülich* (FZJ) started the development of a publication database that went into production in 2002. Later *Deutsches Elektronen-Synchrotron* (DESY) partnered with FZJ and adopted this system, where functionalities required for an institutional Open Access repository were added. While being successfully used for their purposes, around 2009 it became clear, that these databases required substantial functional enhancements. It turned out quickly however, that both systems had diverged significantly over time, and that additions would now require individual development at each partner. This lead to the decision to replace these custom built solutions by some well-established Open Source system and to address local requirements by more general concepts allowing the partners to rely on a common code base, leading to the establishment of the JOIN² initiative[1]: a joint effort tackling the needs of different research centres in a structured, coordinated way.

From user requests it also became clear that the new system should not only be a tool for publication reporting, but it should also

- further the visibility of research done on site. Thus integration into the existing web presence, ensuring high quality indexing by search engines and providing data directly to external services like OpenAIRE (the European OpenScience repository[2]) or BASE (a multidisciplinary academic search engine[3]) became more of a focus.
- allow to derive personalized publications lists easily e.g. for a vita, as well as aggregations for project or institutional pages on the web or via exports for basically any kind of use cases.
- take the requirements of OpenScience into account. This called for a tight integration of the publications database and Open Access repository and for good integration with workflows for both Green and Gold Open Access publication including fee and embargo handling.
- be used as a central document repository not only for own publications, but also for all kinds of works required to conduct the own research. This common request by users requires a good integration with tools for reference formatting like BibTeX or commercial tools for use with word processors as well as good import functions for fast addition of content.
- be used as a central database to report on publications made at the institution. This adds very high demands on data quality and normalization.

These requirements led JOIN²'s search for a suitable software towards repository systems that can handle a data model beyond simple *Dublin Core*[1–4] and are able to cope with large numbers of records while providing fast searches and high availability. It also led to the decision to base JOIN² on individual, closed collections reflecting the organizational structure of the hosting sites. At it's core it is more a web based literature management system. It also demanded for handling of full text files beyond pure Open Access including fine grained access control mechanisms, and while still focusing on publications, it should allow for full text formats beyond PDF.

2. Implementation

After a detailed market analysis, the two institutions decided to go for Invenio[4][5]. Even though it was clear from the start that some additions would have been necessary, it was concluded that the flexible, in libraries well established and understood, meta data model Marc21 is suitable to handle different requirements while keeping a common code base. This was, and still is, considered to be the decisive advantage of Invenio compared to other solutions available. However, some additional development still needed to be done.
The main missing functionalities were identified as:

[1] https://join2.de

[2] https://openaire.eu

[3] https://base-search.net

[4] https://invenio-software.org/legacy

- Web forms, that allow non-experts to submit their publications easily, while ensuring very high meta data quality without much manual curation. This is tackled by importers from commonly used systems[5] to streamline submissions in conjunction with extensive authority control.
- Export interfaces that allow for re-use in bibliographic software[6] and especially on the web.
- Authority control to avoid inconsistent inputs and thus to increase the quality of bibliographic data. This also includes author disambiguation required for personalized publication lists on the web.
- A workflow that allows for at least three steps to guarantee bibliographic quality and proper permission handling for bibliographic data and full text files on the repository side.

It was further agreed that all functions that are offered to the user will be used by staff members as well, and that, by design, there will be no additional hidden functions only visible to staff. This ensures on one hand that everything is working as expected, while on the other hand staff members will always see the same information and in the same way as any user and can thus assist easily if needs be. It also helped to identify and fix weaknesses in the overall submission process and streamline it as much as possible. As a result, e.g. all bibliographic reports are produced using the provided search functionalities and export formats.

From the very beginning GSI Helmholtzzentrum für Schwerionenforschung (GSI) showed interest in the project and joined right before the implementation, as did RWTH Aachen University Library (UB RWTH) only shortly afterwards. DESY and GSI also planned to migrate their library systems to Invenio and, ideally, run them as an integrated system similar to CDS[7] at CERN.

After some preliminary work in 2010, implementation focussing on the publications database began with these four partners in 2011[6].

At the end of 2012 JOIN² reached the first major mile-stone when JuSER went into production to replace the old database *VDB* at Forschungszentrum Jülich. By 2013, all other partners had systems running on site and most were already in production[7]. During 2013 JOIN² also organized the *Invenio User Group Workshop* (IUGW) at *Forschungszentrum Jülich*[8] The *Heinz Meier-Leibnitz Zentrum* (MLZ) in Garching joined as a new partner and their system iMPULSE went into production already in 2014.

Having all sites of the partners online was the next decisive mile-stone for JOIN². Given the diversity of the institutions involved it proved the flexibility and also the viability of the solution. Having a university on board JOIN² had to cope with publications from all fields, ranging from Arts and Humanities to all areas of the Sciences. While this did not allow for any subject specific short cuts[9], it ensured the capability to handle very general use cases and finally paid off as more partners from various fields joined.

Designed as a distributed system from the start, a lot of the infra structure was built and established e.g. for joint development or system roll out. Unlike the usual approach in Invenio, JOIN² always featured a fully automatic system for setup and configuration where the parameters for the latter are organized in `git` repositories. It thus was the first Invenio system that employed an overlay structure as is now common practice. This required quite some work up front, but it proved a necessity as the project expanded later and it allows the project to establish a monthly almost seamless roll out routine. It is also one of the building blocks for the successful `dockerization` of the project, again a technology that barely existed at the start of JOIN².

Having all original partners' instances in production freed up some resources and finally allowed for new partners to come on board. By 2014 the *German Cancer Research Centre* (DFKZ) joined the project[8] while the *German Centre for Neurodegenerative Diseases* (DZNE) already showed first interest as well[9].

[5] doi, pubmed, arXiv, inspire etc.

[6] This required some enhancement of existing exports, especially for EndNote/RIS.

[7] https://cds.cern.ch

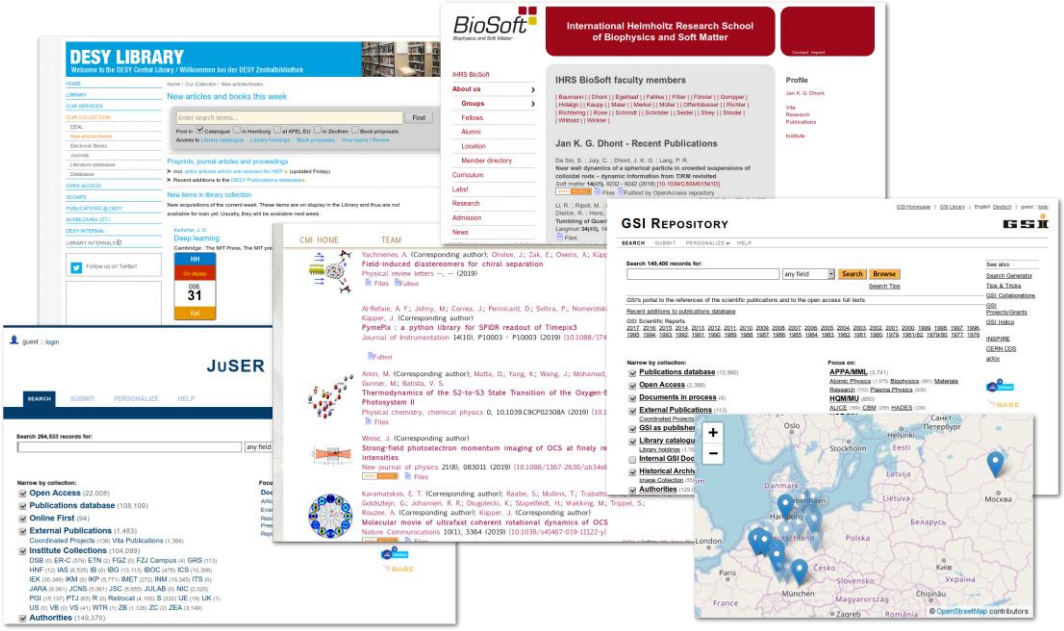
[8] sic!Jülich
https://www.fz-juelich.de/zb/EN/About_us/conferences/%5Bsic!%5D2013/sic_2013_node.html.

[9] Effectivley, it even required to straighten some of those taken by CERN.

Being on the agenda already from the start DESY finally started to migrate their library system (ILS) and integrate it into pubdb, which thus became a central hub for all publication related services at DESY. The most visible part of this integration is the library catalogue. Development on this part was completed in 2017[10] and shortly afterwards GSI was also able to adopt the new functionalities. However, an ILS has quite a few requirements based on workflows in the background reflecting different processes and needs and again the initial approach proved flexible enough to tackle them smoothly.

One of the major issues JOIN² had to solve from the start was name forms in non-English languages, especially for author disambiguation and normalization[11]. In general each name has an *official form* and can have *n* alias forms associated. While in western-European languages for names this involves only a few characters in Latin script, the issue becomes more pronounced in non-Latin scripts like Cyrillic. Thus first contacts with the Joint Institute for Nuclear Research (JINR) date back to the IUGW 2013, mainly due to this common issue.

Again authority records are employed to solve this problem. Here they model people[10], and in order to describe them properly these records also hold additional attributes like an address or multiple name forms. Authority records have been a common concept in the library world for decades[11], while it is a quite new concept to the repository world and did not exist in Invenio. Being a basic requirement however, ideas and work on how to do this in JOIN² date back to the initial project layout. While the implementation was done well before the advent of ORCiD[12], the schema is flexible enough that JOIN² is, and has always been, ready for ORCiD. Effectively, it can even use ORCiDs as primary author identifier at any point in time. The same is true for other kinds of identifiers.


DESY LIBRARY
BioSoft
International Helmholtz Research School of Biophysics and Soft Matter
IHRS BioSoft faculty members
Jan K. G. Dhont - Recent Publications
GSI Repository
JuSER
CMI HOME
TEAM

Sample of screen shots from various instances and their seamless output to the institutes web pages. The map shows the partner institutions of JOIN². Dark colour indicates sites that host the actual systems, light shade refers to individual lab sites associated to one partner. E.g. DESY PUBDB serves the two lab locations of DESY: Hamburg (dark) and Zeuthen (light).
See also the project web site at https://join2.de

After some initial trials to flesh out the functionality of JOIN² for re-use in JDS[13], JINR decided to become a partner in 2017[12] and finally a *Memorandum of Understanding* between JINR

[10] JOIN² features a similar handling for institutes, groups, grants, experiments, journals etc.

[11] see [11] and https://www.loc.gov/marc/authority/ for a full description of Marc Authority

[12] https://orcid.org

[13] http://jds.jinr.ru

and DESY on behalf of JOIN² was signed. The current version of JOIN² now solved the issues with multiple scripts for searching and also displaying the authority records, but there still remain some issues to be solved with regard to multi-script submission. In practice, adding another script results in at least two *official forms* of a name, and the one to be choosen may depend on the language of the data. Shortly after JINR also DNZE joined the project and started implementation in 2018. These two new partners also required some internationalization within the project. While for quite some time the main language (except for code and technical issue tracking) was German, it now moved to English in almost all areas. This also required a rewrite of the documentation, which is mainly tackled by DZNE along with their implementation of DZNEPUB. Furthermore, DZNE was the first partner that employed `docker` from the start, another important mile-stone for the project which implemented an enabling technology for the future[14].

Today, JOIN² operates seven systems in production with two more (DZNE and JINR) to come. More than 28.000 staff members and more than 6.000 visiting scientists have full access to functionalities. JOIN²-systems serve more than 475.000 records out of which more than 68.000 offer a freely accessible full text and 135.000 are shared authorities.

3. Advanced functions

The powerful author disambiguation together with the possibility to store all kinds of publications from whatever time period allow to keep the whole academic record. Using the web export will guarantee pages to stay updated and the reuse of data required for reporting avoids multiple submissions of the same data. This is particularly useful in case of publications with several several authors from the same institute. The integrated full text repository also allows transparent access to the publications for co-workers. In case of Open Access publications the whole community can access them easily from the authors' web pages. Invisible, semantic mark-up is employed for the export to ensure proper indexing by search engines and thus further visibility, while integration in the overall web presence of the research institution gives quite a boost in ranking. The latter again profits from an Open Access full text attached.

Additionally, the JOIN² workflow ensures that it is always safe for authors to attach their articles. No file will be released to Open Access without manual checking for legal restrictions and if necessary, embargoes are handled automatically. This also allows to build an archive of the own achievements over time while adding relevant publications to the institute's collection will build a valuable resource for future work that allows the whole group on campus easy access.

As outlined above even though right now each JOIN²-instance employs a local author identification scheme, the implementation is done such that it already integrates as much as possible with ORCiD. If JOIN² is used as an OpenAccess publishing platform, minting of DOIs can be automatized and will also pass on the ORCiD of the authors and thus can even update the authors' ORCiD profiles automatically if configured to in ORCiD. Alongside ORCiD all author records can also feature a number of other author identifiers. E.g. if the inspire-id is associated with an author even the import of large collaboration papers from INSPIRE-HEP[15] including author association is fast and easy.[16] The main bottle neck is currently parsing and handling web service returns with way beyond 2.000 individual authors, which takes a few seconds.

Recent enhancements of JOIN² are targeted to streamline the publication workflow. The first step optimized the delivery process of theses in High Energy Physics to INSPIRE-HEP. To this end a harvesting routine was established that adds relevant records published by DESYs publishing house on PUBDB to INSPIRE-HEP automatically and in a timely manner. In a second step workflows used in the publishing house for the production of proceedings were optimized. In this case the publishing house not only produces the full volume, but also each contribution as an individual article. The new

[14] RWTH publications, while starting out on a dedicated machine is also already in production running in docker and development using docker is now commonly adopted.

[15] https://inspirehep.net

[16] A close collaboration between INSPIRE-HEP and JOIN² ensures to fetch inspire-ids and provide them for all partners for almost automatic addition to the authors records.

process now ensures that they are grouped together and that all records are interlinked so it becomes easy to navigate trough the volume.[17] Furthermore, the library is now able to just use the very same record for the printed edition of the book as for the digital edition while in the past a second record had to be created. Finally, the same process that was employed to add relevant theses to INSPIRE-HEP now also fetches proceedings including their contributions. For in-house authors the new procedures also ensure proper reporting of the publications and notification of the groups involved.

Another major enhancement was required to handle the current change in the publishing industry. The traditional publication model was based on subscription fees, that is one *pays to read*. With the movement towards full Open Access commercial publishers and societies change more and more to models based on Article Processing Charges (APC). In these models one *pays to publish*.[18] This is becoming an increasingly important part, as can be seen e.g. from the numbers published by the openAPC[19] project which is tracking these expenses internationally. To streamline the internal processes in these cases and also to make the cost aspect transparent for the users new functions were added to the systems. As similar requirements already exist to run the library system in case of book purchases it was possible to employ synergies between both modules. Currently, the definition of a common, xml-based export format for the price information is under discussion with openAPC.

4. Conclusion

Building on a 100% open source framework and around the users' needs, JOIN² has developed a publication database, repository and integrated library system able to address the needs of an expanding set of diverse research centres providing rich functionalities in the simplest way. The definition and enforcement of a uniform publication workflow is at the core of the JOIN² approach. We believe the JOIN² collaboration model to have proven very successful.

Bibliography

[1] Information and documentation — the dublin core metadata element set — part 1: Core elements. (International Organization for Standardization, 2017).

[2] Information and documentation — the dublin core metadata element set — part 2: DCMI properties and classes. (International Organization for Standardization, 2017).

[3] The dublin core metadata element set. (National Information Standards Organization, 2013).

[4] Encoding dublin core metadata in html. (Internet Engineering Task Force, 2010).

[5] Wagner, A. Veröffentlichungsdatenbank und Volltextrepositorium. BIT online 14, 45–48 (2011).

[6] Wagner, A. Ein neues JUWEL? - Publikationsmanagement für Wissenschaft und Administration. in 13 (Spezialbibliotheken - Freund und Follower der Wissenschaft, Jülich (Germany), 9 Nov 2011 - 11 Nov 2011; Forschungszentrum Jülich, Verlag, 2011). doi:10.3204/PUBDB-2017-00882

[7] Wagner, A. Invenio@HGF – status and perspectives. in (Forschungszentrum Jülich; 2nd Invenio User Group Workshop, Jülich (Germany), 18 Nov 2013 - 20 Nov 2013, 2013). doi:10.3204/PUBDB-2017-00869

[8] Wagner, A. & Thiele, R. Invenio@HGF – Collaborative repository infrastructure. in (Open Repositories 2014, Helsinki (Finland), 8 Jun 2014 - 13 Jun 2014, 2014). doi:10.3204/DESY-2014-02793

[17] See e.g. https://doi.org/10.3204/DESY-PROC-2013-04.

[18] However, often unnoticed, even traditional publication models featured quite a number of additional fees like page or colour charges or fees for hybrid Open Access.

[19] https://www.intact-project.org/openapc

[9] Wagner, A. JOIN² – going for the ². in (Deutsches Zentrum für Neurodegenerative Erkrankungen, 2015). doi:10.3204/PUBDB-2015-03786

[10] Wagner, A. Invenio as a library system. in (Heinz Maier-Leibnitz Zentrum; 4th Invenio User Group Workshop, Garching (Germany), 21 Mar 2017 - 24 Mar 2017, 2017). doi:10.3204/PUBDB-2017-01357

[11] Wagner, A. Authority Control in Invenio. in (Forschungszentrum Jülich; 2nd Invenio User Group Workshop, Jülich (Germany), 18 Nov 2013 - 20 Nov 2013, 2013). doi:10.3204/PUBDB-2017-00870

[12] Wagner, A. JOIN² – A scientists toolbox. in (Joint Institute for Nuclear Research, 2018). doi:10.3204/PUBDB-2018-00618