Invenio User Group Workshop 2017

Tuesday 21 March 2017 - Friday 24 March 2017

Heinz Maier-Leibnitz Zentrum (MLZ)

Book of Abstracts

Contents

Legacy / 1

Things you can do dumping your Invenio database into a flat file

Author: Ferran Jorba[1]

[1] *Universitat Autònoma de Barcelona*

Corresponding Author: ferran.jorba@uab.cat

Invenio database design and interfaces are optimized for fast end user search and retrieval. As administrators, we can add indexes at will and use them via web or API. However, many maintenance tasks are not well covered with those indexes.

For most of those cases, reading the records sequentialy is the optimal solution. However, if the database is large enough, reading them via Invenio API may take hours, while the system slows down and it may become unresponsive.

In this presentation I'll show a small Python tool that uses Invenio API and a SQLite database as cache to keep an up to date flat file with your bibliographic records.

We'll see how whith this flat file it is much faster and easier to do tasks like generate specialised statistics, quality control, automatic record enrichment or cleaning, or even creating exotic indexes or counters.

Workshop / 2

A new record editor for Invenio 3

Author: Javier Martin Montull[1]

[1] *CERN*

Corresponding Author: javier.martin.montull@cern.ch

On this presentation, a new record editor will be presented. Current version under development can be found in https://github.com/inveniosoftware-contrib/ng2-json-editor. This editor uses JSON as its native data format, provides many configuration options and can handle very large JSON documents. An update on the development status and pointers to how to use it in your own installation will be provided.

Services round table / 3

EUDAT B2SHARE overview

Author: Nicolas Harraudeau[1]

[1] *CERN*

Corresponding Author: nicolas.harraudeau@cern.ch

B2SHARE is a digital library for research data software provided by EUDAT. It has recently been migrated to Invenio 3 and put in production December 2016.
This presentation will show a brief overview of its features and how it integrates with other EUDAT services.

Research data / 4

Dynamic metadata model for B2Share

Author: Nicolas Harraudeau[1]

[1] *CERN*

Corresponding Author: nicolas.harraudeau@cern.ch

Invenio 3 validates metadata format using JSON Schemas. This presentation will show how B2Share enables its users to create their own custom schemas and share them with other communities.

Services round table / 5

Invenio @ JINR

Authors: Tatiana Zaikina[1]; Genis Musulmanbekov[2]; Irina Filozova[3]; Roman Semenov[4]

[1] *Joint Institute for Nuclear Research*

[2] *JINR*

[3] *Joint Inst. for Nuclear Research (RU)*

[4] *Joint Institute for Nuclear Research (RU)*

Corresponding Authors: roman.semenov@cern.ch, ztanya@jinr.ru, genis@jinr.ru, irina.filozova@cern.ch

JINR open access repository, JDS (JINR Document Server), launched on the Invenio platform is functioning since 2009. Started with Invenio v.99 and now updated it to v.1.2.2.
JDS collections include published articles, books, theses, conference proceedings, audio, video materials, etc. Various methods of ingesting of documents into JDS and updating its content are applied: submission by authors, harvesting, (automatic) uploading. Further development of JDS is connected with the project "JINR corporate information system" aimed as information support of scientific researches performed at JINR. Within the project we are creating a collection "Authority" which is intended to be a core of this system.

Workshop / 6

Machine Learning examples on Invenio

Author: Samuele Kaplun[1]

[1] *CERN*

Corresponding Author: samuele.kaplun@cern.ch

This talk will present the different Machine Learning tools that the INSPIRE is developing and integrating in order to automatize as much as possible content selection and curation in a subject based repository.

7

Troubleshooting

8

Hands-on: Migration dumping data

9

Troubleshooting

10

Troubleshooting

11

Workshop Dinner @ Gasthof Neuwirt

12

Tour

13

Troubleshooting

Workshop / 14

Welcome

Corresponding Author: connie.hesse@frm2.tum.de

Workshop / 15

Overview and Goals

Corresponding Author: jose.benito.gonzalez@cern.ch

Workshop / 16

Invenio: State of the Union

Corresponding Author: lars.holm.nielsen@cern.ch

Workshop / 17

Community & Processes

Workshop / 18

Feedback on workshop

Workshop / 19

Workshop summary

Workshop / 20

Feedback on Invenio Future Directions

Workshop / 21

Invenio v3 Overview

Corresponding Author: lars.holm.nielsen@cern.ch

Hands-on (Develop v3) / 22

Architecture and module overview

Corresponding Author: lars.holm.nielsen@cern.ch

Services round table / 23

Invenio @ ….

Services round table / 24

Invenio @ …

Services round table / 25

Invenio @ …

Services round table / 26

Invenio @ JOIN2

Corresponding Author: alexander.wagner@desy.de

Services round table / 27

Invenio @ IAEA

Corresponding Author: j.garcia-llopis@iaea.org

Services round table / 28

Invenio @ Universitat Autònoma de Barcelona

Corresponding Author: ferran.jorba@uab.cat

Services round table / 29

Invenio ILS as SaaS @ TIND

Corresponding Author: audun@tind.io

Services round table / **30**

Invenio @ RWTH Aachen University

Services round table / **31**

Invenio @ RERO

Corresponding Author: johnny.mariethoz@rero.ch

Services round table / **32**

Invenio @ CERN Scientific Information Services (INSPIRE / HEP-Data / SCOAP3)

Corresponding Author: samuele.kaplun@cern.ch

Services round table / **33**

Invenio @ CERN IT (CDS, B2SHARE, Zenodo, OpenData, Analysis Preservation, OAIS Archival Store)

Corresponding Author: jose.benito.gonzalez@cern.ch

Research data / **34**

Invenio as one Module within a Holistic Service Suite for Research Data Management

Author: Dominik Schmitz[1]

[1] *RWTH Aachen University*

Corresponding Author: d.schmitz@ub.rwth-aachen.de

Research data management is a duty a university or research institute can not ignore any longer. But setting up a suitable infrastructure is cumbersome and ill-supported by national or international infrastructures yet, in particular in Germany [1]. At the same time monolithic IT solutions encompassing the whole data lifecycle as well as the entire university or research institute are not an option since there is much too much development and there are far too many changes and disciplines involved, in particular when looking into solutions that really support individual research units.
There are some prominent projects, mainly ZENODO (http://zenodo.org) and EUDAT (http://eudat.eu), funded by the EU that make use of the Invenio framework mainly for publishing research data.
Yet publishing is only one component of research data management. How about keeping data not be published, long-term preservation, or linking publications to its foundational data? Various different approaches and tools support different aspects of research data management and need to be combined into a holistic and adaptable service suite.
This presentation shows how RWTH Aachen University makes use of the Invenio and in particular

the JOIN2 infrastructure as a module within this service suite. DOI minting, linkage between data records and towards authority files for people, institutes, and projects, and alternative storage facilities are some of the topics that will be addressed. Overall, we point out current achievements as well as open challenges.

[1] Leistung aus Vielfalt: Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland, Göttingen : Rat für Informationsinfrastrukturen, URN: urn:nbn:de:101:1-201606229098, 2016

Legacy / 35

What is needed for effective open access workflows?

Author: Claudia Frick[1]

Co-author: join2

[1] *Forschungszentrum Juelich*

Corresponding Author: c.frick@fz-juelich.de

Institutions and funders are pushing forward open access with ever new guidelines and policies. Since institutional repositories are important maintainers of green open access, they should support easy and fast workflows for researchers and libraries to release publications. Based on the requirements specification of researchers, libraries and publishers, possible supporting software extensions are discussed. How does a typical workflow look like? What has to be considered by the researchers and by the editors in the library before releasing a green open access publication? Where and how can software support and improve existing workflows?

Legacy / 36

Matching and merging

Author: Kirsten Sachs[1]

[1] *DESY*

Corresponding Author: kirsten.sachs@desy.de

When harvesting information from different sources it is necessary to identify identical objects. If both have the same unique identifier like a DOI or a report-number this is trivial but unfortunately a rare case.

Most of the time matching is mainly based on author and title information. However, titles may change significantly from preprint to publication and depending on the type of the publication (journal paper, conference contribution, thesis) even identical basic metadata would lead to separate records.

In general a two-step process is needed:
a) search for potential candidates. Here it is necessary to define a search query with a high efficiency. However, if the search is too fuzzy, the number of records as search result is too large and matching becomes not feasible. Restriction to a limited scope of records is helpful.
b) confirmation of the match. Depending on the strategy clear results can be treated automatically, whereas doubtful cases might be presented to a human for final decision. In both cases it is essential to have enough information.

For a reliable match good quality of uniform metadata is essential and in many cases processing of content information like abstract, references or fulltext is needed.

Once two records have been identified as equal or existing information receives an update, the information needs to be merged. There are obvious cases where one source always supersedes another, maybe some information comes only from one source. But to add e.g. an ORCID from one source to the author and affiliation from another source requires the identification of corresponding information.

Experience from INSPIRE shows what is currently done (fields with controlled vocabulary), what is doable (fields where the content can be identified) and where merging is not feasible but one version simply overwrites another.

What can be done automatically, which tools are needed, when is human intervention necessary? When is it worthwhile to overwrite (i.e. delete) manually curated, high quality information?

Legacy / 37

Invenio as a library system

Author: Alexander Wagner[1]

[1] *Deutsches Elektronensynchrotron DESY, Hamburg*

Corresponding Author: alexander.wagner@desy.de

As a join2 partner, DESY library uses Invenio already for it's publication database and institutional repository. The next logical step is to also migrate the library catalogue from the currently used Aleph system to Invenio. Starting out with a short introduction of how to migrate Aleph. This includes the migration of bibliographic data as well as holdings but also movement data, current loans etc.

The talk also outlines some of the new additions required to run Invenio as an ILS at DESY based on the infrastructure already existing. E.g. it is necessary for DESY to interact with RFID based self service terminals, barcode based library cards and external patrons who have not DESY account etc.

Legacy / 38

Article Processing Charges and OpenAPC

Author: Alexander Wagner[1]

[1] *Deutsches Elektronensynchrotron DESY, Hamburg*

Corresponding Author: alexander.wagner@desy.de

The publication landscape is about to change. While being largely operated by subscription based journals in the past, recent political decisions force the publishing industry towards OpenAccess. Especially, the publication of the Finch report in 2012 put APC based Gold OpenAccess models almost everywhere on the agenda. These models also require quite some adoptions for library work flows to handle payments, bills and centralized funds for publication fees. Sometimes handled in specialized systems (e.g. first setups in Jülich) pretty early on discussions started to handle APCs in local repositories which would also hold the OpenAccess content resulting from these fees, e.g. the University of Regenburg uses ePrints for this purpose.

Backed up by the OpenData movmement, libraries also saw opportunity to exchange data about fees payed. Thus, OpenAPC.de was born in 2014 on github to facilitate this exchange and aggregate large amounts of data for evaluation and comparison. Using the repository to hold payment data usage of OAI-PMH is immediate. Thus, join2 and the University of Regensburg developed an interchange format for APC data that allows easy and automatic delivery to OpenAPC.

This talk outlines a working solution for APC management and hook up with OpenAPC based on Invenio as implemented in join2.

Services round table / 39

Invenio 3 - Call to Action

Author: Alexander Wagner[1]

[1] *Deutsches Elektronensynchrotron DESY, Hamburg*

Corresponding Author: alexander.wagner@desy.de

As a very flexbile system Invenio is also one of the most complex repository solution available today. This is especially true for the upcoming Invenio 3, which delivers more or less only the building blocks for a repository solution than a ready to run application but that are adoptable to almost any use case. On the other hand we see a huge convergence in repository usage if some concepts are tackled in a more abstract manner, usually on the data side instead of the software side.

The experiences in the join2 project showed us that while starting out as individual solutions with a host of configurability in mind, over time our systems more and more converged on the basis of common concepts that now run successfully at a number of quite different institutions.

While the famous *Atlantis Institute of Fictive Science* was more or less a showcase of Invenios capabilities join2 now want's to rise the question to the community:

Isn't it time to set up a real world usable version of Atlantis?

A version of Atlantis could be a generally usable repository based on Invenio that can be run everywhere requiring *configuration* instead of programming.

Legacy / 40

The usages of JOIN2 authority records

Authors: Robert Thiele[1]; Katrin Grosse[2]

[1] *Deutsches Elektronen-Synchrotron DESY*

[2] *GSI Helmholtzzentrum für Schwerionenforschung GmbH*

Corresponding Authors: robert.thiele@desy.de, k.grosse@gsi.de

An important base of the common JOIN2 repository infrastructure of DESY, DKFZ, FZJ, GSI, MLZ and RWTH Aachen are about 134 000 authority records for grants, projects, large-scale infrastructures, cooperations, journals, and different kinds of keys. All instances are using the authorities together.
We will present how these authority data are used for different purposes e.g. the recent and upcoming obligations to report to regard to our funding and the data export to openAire. Furthermore, we discuss this in dependence to the German "Kerndatensatz Forschung", which will be the new standard for future.

Legacy / 41

ORCID implementation in Invenio 1.1

Author: Torsten Bronger[1]

[1] *Forschungzentrum Jülich*

Corresponding Author: t.bronger@fz-juelich.de

We present an extension to the Invenio 1.1 software for semi-automatically harvesting ORCID IDs of users and allowing them to upload publications to their respective ORCID profile. This extension was created in the context of the Join2 initiative, however, it can easily be adapted to other Invenio instances because it is only loosely coupled with Invenio itself. It opens its own local webserver to handle the additional endpoints, and calls Invenio API functions and command line programs to interact with the database. We also present a recommended workflow for successfully harvesting ORCID ID in an institution. The implementation is realised in well-documented Python 2.6 and Go and will be published as Free Software.

Workshop / 42

Invenio 3 - Call for action

Corresponding Author: alexander.wagner@desy.de

Hands-on (Getting started) / 43

Getting started with v3

Corresponding Author: tibor.simko@cern.ch

Hands-on (Getting started) / 44

End-user tour of v3

Corresponding Author: tibor.simko@cern.ch

Hands-on (Getting started) / 45

Loading records, indexing, REST API, OAI-PMH server.

Corresponding Author: tibor.simko@cern.ch

Legacy / 46

Migrating records from v1.2 to v3

Corresponding Author: esteban.jose.garcia.gabancho@cern.ch

Legacy / 47

INSPIRE live migration

Corresponding Author: samuele.kaplun@cern.ch

Legacy / 48

CERN Document Server migration of 1.2M records

Corresponding Author: ludmila.marian@cern.ch

Legacy / 49

TIND Integrated Library System

Corresponding Author: audun@tind.io

Hands-on (Customisations) / 50

Simple customisations of logo, facets, sort options, query parser and record templating

Corresponding Authors: charalampos.tzovanakis@cern.ch, javier.martin.montull@cern.ch

Hands-on (Customisations) / 51

Tutorial/tour: v3 data model and indexing

Author: Nicolas Harraudeau[1]

Co-author: Lars Holm Nielsen [1]

[1] *CERN*

Corresponding Authors: nicolas.harraudeau@cern.ch, lars.holm.nielsen@cern.ch

Hands-on (Customisations) / 52

Tutorial: Enabling ORCID login in V3

Corresponding Author: samuele.kaplun@cern.ch

Service management / 53

Services for Invenio v3 (Elasticsearch, PostgreSQL, …)

Corresponding Author: guillaume.lastecoueres@cern.ch

Service management / 54

Monitoring your V3 infrastructure

Corresponding Author: lars.holm.nielsen@cern.ch

Service management / 55

High availability for Invenio v3

Corresponding Author: esteban.jose.garcia.gabancho@cern.ch

Service management / 56

Deploying Invenio v3

Corresponding Author: esteban.jose.garcia.gabancho@cern.ch

Service management / 57

Deploying with Docker

Corresponding Author: audun@tind.io

Research data / 58

Caltech RDM by TIND

Corresponding Author: audun@tind.io

Research data / 59

Two Petabytes in Invenio? CERN (Open) Data

Corresponding Author: tibor.simko@cern.ch

Research data / 60

Handling large files and versioning data sets

Corresponding Author: lars.holm.nielsen@cern.ch

Workshop / 61

CERN Document Server Videos

Corresponding Author: ludmila.marian@cern.ch

Hands-on (Develop v3) / 62

Develop a simple v3 module

Corresponding Authors: remi.ducceschi@cern.ch, charalampos.tzovanakis@cern.ch

Hands-on (Develop v3) / 63

Package your Invenio v3 module

Corresponding Authors: remi.ducceschi@cern.ch, charalampos.tzovanakis@cern.ch

Community / 64

Invenio community processes

Corresponding Author: lars.holm.nielsen@cern.ch

Community / 65

Hands-on: Translating Invenio v1 and v3

Corresponding Author: tibor.simko@cern.ch

Community / 66

Group Brainstorming: Community

Corresponding Authors: alexander.wagner@desy.de, jose.benito.gonzalez@cern.ch

Community / 67

Group Brainstorming: Community

Corresponding Authors: jose.benito.gonzalez@cern.ch, alexander.wagner@desy.de

Community / 68

Feedback on workshop

Corresponding Author: lars.holm.nielsen@cern.ch

Community / 69

Workshop summary and closing

Corresponding Author: connie.hesse@frm2.tum.de

Workshop / 70

CERN Archival Store: Invenio Archivematica integration

Corresponding Author: remi.ducceschi@cern.ch