INVENIO INTEREST GROUP PROPOSAL

OPEN REPOSITORIES 2014

Samuele Kaplun (CERN) and Wojciech Ziolek (CERN)

Keywords: aggregation; normalization; BibCheck; BibTasklet;

Invenio as a platform to implement the SCOAP³ Repository

Abstract

In the context of the SCOAP³ project, the SCOAP³ Repository had to be set up in order to aggregate in a central place all the SCOAP³ publications from 10 different journals in high energy physics. This has been accomplished by agreeing with publishers how data should reach the Repository (formats, packaging, transport protocols), how metadata has to be normalized, and how this data is finally made available to the user.

It will be presented how the Invenio Digital Library has been adapted in order to meet the requirements of SCOAP³ (namely, rather than being oriented to researchers, the repository has to be an administrative tool, and a gateway towards other services), how core modules of Invenio (OAIRepository, BibCheck, BibFormat, BibTasklet) made it possible to adapt to the several specificities of each publisher, and how, by carefully designing workflows, it has been possible to implement an agile and quick development that allowed us to respect the tight deadlines of the project.

Ingestion of data into the repository

Robotupload

Two participating publishers in SCOAP³ agreed to directly push articles to the repository using Invenio **Robotupload API**, with metadata already represented in MARCXML and with Invenio's custom FFT fields to fetch the actual articles. The configuration of this setup will be presented, in particular how to carefully allow publishers to submit and modify only the records they are authorized to.

OAI-PMH

One publisher is providing data in MARCXML via OAI-PMH. It will be presented how a **filter** has been set up to carefully select what operations to perform with incoming records.

FTP servers

Three publishers are providing data by depositing .zip packages in their local FTP servers to which the SCOAP³ Repository has been granted access. It will be presented how the

BibTasklet framework allowed us to implement, in an agile way, scripts that were able to understand these packages, and how these have been transformed into workflows that are regularly executed as any other Invenio bibliographic task.

Automatic metadata normalization

Since articles are coming from 11 publishers in several different ways, metadata are not always consistent in the first instance. It will be presented how the newly integrated **BibCheck** module has been exploited by writing several plugins and rules that automatically adjust the incoming metadata (e.g. for proper journal name normalization, for extracting a normalized country name from affiliation strings, for proper license information…)

Automatic metadata enrichment

Of great importance to the SCOAP^3 compliance for articles is that these had to reach the Repository not later than 24 hours from their appearance online. To this aim, an ad-hoc module has been introduced into SCOAP^3 Invenio overlay to query CrossRef/FundRef APIs and discover DOIs as they are registered by each publisher. A bibtasklet has been introduced to harvest this information. A BibCheck plugin has been added to enrich records with DOI timings.

Usability and User experience

Being not oriented to researchers rather to libraries and third party services, any user-oriented non-necessary Invenio functionality has been disabled. This includes authentication, baskets, alerts, submissions, circulation… it is presented how this can be achieved in Invenio stable by tuning the **WebStyle** module and by curating authorizations in **WebAccess**. Additionally it will be presented how **MathJax** integration into Invenio has been amended to support both MathML and LaTeX markup.

Search Engine Optimizations

Articles reaching the SCOAP^3 Repository need to be exposed to search engines. It will be presented how, by following native Invenio conventions, data is easily made available to e.g. Google Scholar and how sitemaps are generated for optimal site crawling by using **BibExport**.

Conclusions

Best practices for Invenio configuration and extension and lessons learnt are presented, in the context of a tight deadline and clear project requirements for setting up the SCOAP^3 Repository.