Invenio:
A Modern Digital Library for Grey Literature

Jérôme Caffaro, CERN Samuele Kaplun, CERN

November 25, 2010

Abstract

Grey literature has historically played a key role for researchers in the field of High-Energy Physics (HEP). Consequently CERN (European Organization for Nuclear Research) as the world's largest particle physics laboratory has always been facing the challenge of distributing and archiving grey material. Invenio, an open-source repository software, has been developed as part of CERN's institutional repository strategy to answer these needs.

In this document we describe how the particular context of grey literature within the HEP community shaped the development of Invenio. We focus on the strategies that have been established in order to process grey material within the software and we analyse how it is used in a real production environment, the CERN Document Server (CDS).

1 Introduction

1.1 Background

CERN The European Organization for Nuclear Research in Geneva is the world's largest particle physics laboratory. Originally founded in 1954 by 12 European countries, CERN has established a solid reputation in scientific research throughout history. CERN is currently run by 20 European member states with over 40 additional participating observers (states and organizations). About 8,000 scientists from around 580 universities come to CERN to work on their research. The main current research program at CERN is the LHC (Large Hadron Collider), the largest (a 27km ring of superconducting magnets) and most powerful accelerator, smashing particles together to understand the basic constituent of matter. Over a thousand publications are published yearly by CERN scientists in established journals.

HEP Community The High Energy Physics (HEP) community is estimated to have about 20,000 scientists. It essentially comprises researchers working in the major particle physics laboratories around the world such as CERN, Desy (Germany) Fermilab (USA), SLAC (USA) and KEK (Japan).

1.2 Invenio

Invenio is an integrated digital library system [1] originally developed at CERN to run the CERN Document Server (CDS). It is currently one of the largest institutional repositories worldwide. It was started over 15 years ago and has matured through many release cycles. Invenio is a GPL2 Open Source project based on an Apache/WSGI+Python+MySQL architecture. Its modular design enables it to serve a wide variety of requirements, from a multimedia digital object repository, to a web journal, to a fully functional digital library. The development strategy used to implement Invenio ensures that it is flexible in every layer. Being based on open standards such as MARCXML and OAI-PMH 2.0 its interoperability with other digital libraries is guaranteed. Having been originally designed to cope with the CERN requirements for digital object management, Invenio is suitable for middle-to-large scale digital repositories (100K∼10M records).

2 Grey Material at CERN

Invenio software was born in a rich grey literature producing environment [2]. One early major impact on the HEP community, and by consequence on the development of Invenio was the definition of a strong policy towards the dissemination of the work done at CERN. Indeed the convention that established CERN in 1954 states that "*[...] the results of its experimental and theoretical work shall be published or otherwise made generally available*" [3]. The implied openness of this mission forged a strong idea of responsibility for the community to give access not only to published documents, but also to additional material produced by the organization. CERN being an international organization, involving collaborations with a large number of universities and institutions, efficient sharing of information was a primary concern not only for scientific results, but also for all the material necessary for good coordination and proper running of the experiments: engineering drawings, technical reports, notes, etc. containing important scientific or technical data, but not suitable for publication in journals.

An important trend that took off among HEP researchers more than 50 years ago was the habit of mailing to their peers printed copies of their work at the time of submission to journals [4], reducing by several months the access to results of a possibly major importance in the context of the laboratory: machines and tools built for the CERN experiments being giant and technically advanced prototypes, they require several iterations

of optimizations which can be performed through the analysis of early experimental results. Reducing the duration of these cycles was necessary to help keeping the cost of the experiments as low as possible.
Early experimental results would also be used as input by theoretical physicists in order to refine existing theories or build new ones, and suggest new areas of study for experimental physics [5].

Another early concern expressed by the researchers at CERN was simply related to the physical access to grey literature. It is common for several thousands of scientists and engineers to work on the same experiment due to its complexity and size. A geographical distribution of the experts is inevitable, considering both the country of origin of these experts and location on the experiment(s) site(s) for practical reasons. Coordination of such large projects is only possible through formal, written forms. That particular need of giving access to a large amount of information in an electronic way resulted in the creation of the World Wide Web in 1989, as a proposal by Tim Berners Lee [6].

In the last two decades, HEP continued to pioneer solutions in scholarly communications. SPIRES, the SLAC (Stanford Linear Accelerator) literature database became in 1991 the first web server in the USA and the first online database in the world [7]. The same year arXiv.org (named at the time "LANL preprint archive") opened its web front-end, first as repository of physics preprints before expanding to other fields of science. In 1993 CERN released its preprints database on the web as an early version of the Invenio software, with the initial goal of fulfilling CERN's needs of access to grey literature.

3 Invenio for Grey Literature

In this section we review the strategies adopted in Invenio to support the management of grey literature. Examples of applications to real production systems running Invenio are given. In particular numbers given in the following paragraphs are updated statistics of the CERN Document Server (powered by Invenio) for November 2010.

MARCXML as core metadata format MARCXML is the core bibliographic metadata format of Invenio. It offers all the required flexibility to model a great variety of digital assets.

Being a library standard, MARCXML offers the advantage of being well-known by professional librarians, giving a unique chance for grey material to be curated by the institutional library team. The very same tools used to manage published material can be used to process grey material, ensuring higher quality data and helping grey literature find its way more to standard institutional processes more easily.

Flexible metadata-formatting layer A flexible metadata formatting layer in combination with the MARCXML format allows the visualization of practically any digital asset within Invenio. An accessible HTML-like markup offers the opportunity for librarians to define the display of the managed records.

Additional support for XSLT at the level of the formatting layer enables easy conversion from MARCXML to other XML flavor formats in a standard way.

CDS uses 122 different formatting templates, approximatively half of them being used to prepare search results output, and the other half providing detailed information about the records. The templates deal with standard preprint objects, as well as video, audio or photo content. Combined with a customizable collection tree, it is possible to offer subject-based "portals" regardless of the actual type of contained material.

Customizable workflow engine The submission system of Invenio lets administrators configure their own customized workflows. The framework offers the tools to create web front-ends for users to submit data (metadata and files), and an extensible set of functions to process the collected data.

Typical workflows result in the creation of a new record in the repository. Thanks to the mapping of the collected data to MARCXML, the flexibility offered by the submission system of Invenio regarding the type of supported data can be extended to the archival of this data. Submissions of Invenio can also be the starting point for subject specific jobs such as OCR (Optical Character Recognition) for scanned documents, and image downsizing for the creation of web versions.

91 different workflows are currently maintained on the CERN Document Server, some very similar to the basic workflow described above and other implementing complex reviewing and approval systems implicating thousands of different users. An average of more than 30 documents are submitted per day (less during week-ends) through these web-based workflows (300 documents per day when considering alternative input methods).

Collaborative tools Invenio supports the creation of user groups (local or derived from the institution identity management system) and "baskets" letting users share information in a controlled, targeted manner. This feature is particularly useful in order to provide a community-based selection of unpublished documents at a quality that a ranking algorithm cannot match. Shared baskets can then offer a fast changing community-based hierarchical structure of data that Invenio main navigable collection tree cannot provide. In November 2010, CDS counted 6,245 non-empty baskets set up by 4,575 distinct users, covering about 10% of the whole archive. 6% of the baskets were shared among several users.

Invenio also features basic commenting and reviewing capabilities enabling a better understanding of the quality of the material. Consequently it also archives information about the documents which in the past was wrongly regarded as transient information

only. The most active collection on CDS, regarding the number of comments, gets an average of about 30 comments per day through its built-in commenting system, usually on recent work being under peer-review at CERN.

Access control also plays an important role in helping the ingestion of grey material: providing an adequate, secure and restricted collaborative workspace for draft documents is an incentive for users to move their workflow to a central server, hence giving more opportunity for the final document to be made publicly (or not) available. Indeed Invenio can accompany documents through their life cycle, thanks to the integration of an advanced role-based access control system into the flexible workflow engine.

To bring a better awareness of the quality of the material to users and help with the discovery of documents of possible interest, Invenio display recommendation based on document usage statistics (*"People who viewed this page also viewed..."*) and features download/citation history graphs.

Search engine A major concern for large repositories of grey material is to provide an efficient way to retrieve new and archived material. Invenio includes a very fast search engine optimized for large repositories (millions of documents) on simple infrastructures, combining metadata and fulltext search in a simple Google-like query language. Advanced users are also given the opportunity to perform advanced queries, such as *find document written by Ellis from years 2000 to 2010, mentioning "higgs boson" in the fulltext, referring to documents written by Randall, and cited more than 50 times.* The CERN Document Server serves about 25,000 queries per day, for an archive of about 1 million records.
Retrieved documents can be ranked according to several techniques, such as "word similarity" ranking or "citation-graph" [8] based ranking etc. in order to accommodate to the type of searched material: for example a researcher new to some subject is more likely to search for general reference documents while an expert might rather be looking for all new material in his field of interest.

The Invenio search engine technology is at the core of many functionalities offered by the software. For example combined with the flexible metadata-formatting layer, it can provide personalized search-based RSS feeds or email alerts: new results to some specific search queries such as the sample one mentioned above can be sent periodically to the subscribers. This has proven to be an essential functionality for CERN physicists in need of the latest information on some very specific topics. CDS counts more than 12,000 RSS subscriptions set up by 3000 distinct users (IP-based). 2417 emails alerts have also been set up by 1615 users.

Other use cases of the search engine include the suggestion of documents similar to a given one, or the creation of the bibliography (BibTeX) of a given author, personalized podcasts, etc.

Interoperability Invenio implements standard protocols to help the ingestion and dissemination of documents. Though these protocols are usually independent of document-type, they can still suffer from the conversion process usually occurring to ensure that repositories use a common language. For example OAI-PMH is able to support any document type, but most repositories only support the Dublin Core schema, hence narrowing down the possibility to use this protocol in some scenarios. Invenio is able to export and import any metadata format in OAI-PMH thanks to the underlying layers supporting custom conversion templates. For example OAI-PMH is used at CERN to feed an installation of Invenio with conference and meeting objects coming from the institutional conference management system Indico.

Another example of usage of OAI-PMH in Invenio is in the context of the OPENAIRE project, which is planning to exchange usage statistics among participating repositories through this protocol.

Integrated digital library Invenio is a multi-purpose repository software: not exclusively designed for grey material, it offers the advantage of being a solution for the common needs of a library. It results in lower infrastructure maintenance costs by grouping several library services and processes on a single server. Reusing the same technology and concepts for these different services is also reducing the learning curve to master the necessary tools. It becomes much easier to justify the cost of supporting grey material within the institution.

4 Conclusions

Invenio is a well-established open source repository software. The context in which the software was conceived and then further developed has played an important role in defining a core set of features suitable for the ingestion, processing and distribution of grey material.

The performance and flexibility of the software has led to its adoption in a variety of scenarios, strengthening the will to drive the development efforts towards an increased support for grey material.

References

[1] Invenio official website.
`http://invenio-software.org/`
(Last visited on 19 November 2010)

[2] L. Goldschmidt-Clermont, *Communication Patterns in High-Energy Physics*; High Energy Physics Libraries Webzine, issue 6, March 2002.
`http://library.web.cern.ch/library/Webzine/6/papers/1/`
(Last visited on 18 November 2010)

[3] *Convention for the Establishment of a European Organization for Nuclear Research*, Paris, 1st July, 1953
`http://council.web.cern.ch/council/en/Governance/Convention.html`
(Last visited on 18 November 2010)

[4] Anne-Gentil Beccot et al., *Information Resources in High-Energy Physics: Surveying the Present Landscape and Charting the Future Course*, J.Am.Soc.Inf.Sci.60, 150-160, (2009)

[5] Anne-Gentil Beccot, *How do High Energy Physics scholars search their information?*, Grey J. 4, 1 (2008).

[6] Tim Berners-Lee *Information Management: A Proposal* CERN-DD-89-001-OC, 1989
`http://cdsweb.cern.ch/record/369245`

[7] L. Addis, *Brief and Biased History of Preprint and Database Activities at the SLAC Library, 1962-1994.*
`http://www.slac.stanford.edu/spires/papers/history.html`
(Last visited on 19 November 2010)

[8] Ludmila Marian et al., *Citation graph based ranking in Invenio.* LNCS (Research and Advanced Technology for Digital Libraries) 6273 (2010)