Invenio: A Modern Digital Library System

Introduction

Invenio is an integrated digital library system originally developed at CERN to run the CERN Document Server, currently one of the largest institutional repositories worldwide. It was started over 15 years ago and has been 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 usages, from a multimedia digital object repository, to a web journal, to a fully functional digital library. The development strategy used to implement Invenio ensures it is flexible in any layer. Being based on open standards such as MARCXML and OAI-PMH 2.0 its interoperability with other digital libraries is guaranteed.

Being 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). Records can be of any nature (e.g. papers, books, photos, videos).

This presentation will introduce the different features of Invenio, their usage in the CERN context and how other institutions and projects are also driving some of its development.

Usage

Besides being used to run the CERN Document Server (which is ranked 4th in the Webometrics Top 400 institutional repositories), Invenio has also been chosen by several other big institutions or projects. Among them it is about to be used to serve the SPIRES High Energy Physics information through the recently launched INSPIRE service, that will become the repository of reference for high energy physics. It is used by the EPFL (the Ecole Polytechnique Fédérale de Lausanne, 5th in the Top 400) to power Infoscience, their institutional repository). ADS (the SAO/NASA Astrophysics Data System, 1st in the Top 400) has expressed interest to move to an Invenio-based platform, while starting a collaboration at the data level between astrophysics and high energy physics worlds. For example in Spain Invenio is already used by the Dipòsit Digital de Documents (DDD) (Universitat Autònoma de Barcelona) and by the “Repositorio Digital de la Universidad de Zaragoza” (University of Zaragoza). Recently the European Commission has chosen Invenio to become part of two important projects D4ScienceII, and OpenAIRE, with the latter having the goal to set up a portal where all the EC funded research project documents will be available.

Features

Let us demonstrate the key Invenio features by following the lifetime of a record in the system, from ingestion to dissemination.

As an integrated digital library, Invenio is built to allow the management of digital objects and the meta-data associated with them. Meta-data are stored in the MARCXML format, which has the flexibility to store any kind of information (scientific papers, books, multimedia documents, etc.) and which can be easily converted, thanks to XML style-sheets to any other meta-data format.

Ingesting records

Data and Meta-data can enter into Invenio powered repositories through different ways. They can be manually deposited by an author (or someone delegated to this action), through custom and fully configurable web interfaces, backed by an equally configurable work-flow architecture, that lets the submission administrator set up any kind of procedure triggered by a deposition. Meta-data and data can as well be harvested regularly from OAI-PMH compatible repositories. Data can also be input into the system from any non-standard format by writing ad-hoc conversion scripts either in XSLT or in the proprietary BibConvert's language. Users also have the possibility to upload records by

sending emails that will be parsed automatically. The system can be configured to accept submission by robots thanks to the availability of a simple web API.

Digital object manipulation

Fully fledged tools to integrate and manipulate digital objects are available. In particular Invenio can manage multiple formats for the same document, multiple revisions, automatic conversions between formats (including, out of the box, scanned PDF to PDF with OCRed text, or Microsoft Office conversions, PDF/A...), automatic stamping of PDF documents with watermarks, creating icons, automatic text extraction to allow for full-text searching, automatic meta-data extraction (e.g. from PDF's XMP and JPEG's EXIF), and many more. Data consistency is also ensured by MD5 checksum being continuously verified.

Indexing and ranking

All the new records ingested into the system and their modifications are indexed by a very fast indexing engine that has been developed in-house. All the indexing stages are configurable; which metadata fields to index, how to extract tokens, which language is expected in order to apply stemming and removing stop words, and so on. A plug-in based ranking engine takes care of creating special indexes to be used at search time for sorting results based on word similarity, citations, download statistics, and to present citation networks (e.g. “cited-by”, “co-cited-with”)

Curating records

Curators have powerful web-based tools to manipulate records, whilst maintaining change histories, auto-completion of fields based on fully configurable knowledge bases (supporting RDF ontologies), multi-record modification, assisted duplicate record finding and an efficient AJAX-based record merging tool, and so on.

Curators can also exploit flexible tools to perform meta-data compliance checks against configurable rules.

Automatic tools are also available to extract citations and references from scientific papers and to extract keywords based on a given ontology, to enrich the available meta-data.

Distributed curation activities can be instantiated through integration of Invenio with RT, one of the most famous open source ticketing systems.

Displaying records

Bibliographic records can be displayed in any format, thanks to an easy and fully-configurable formatting layer, which allows creation of record-based rules to choose which representation of a record to use, while each representation can be fine tuned to fit any particular need. Thus any kind of content can be hosted and properly displayed, such as multimedia documents, books, address book contacts, and so on.

Records can be organized in a flexible hierarchical structure using real and virtual collections. Each collection can in turn be customized to choose the look and feel and the kind of features enabled.

The whole web appearance of Invenio can be fully customized thanks to an integrated templating system, thus facilitating integration of Invenio into an existing institutional portal.

Since Invenio is being developed in an international environment it supports internationalization and comes translated already in 25 different languages.

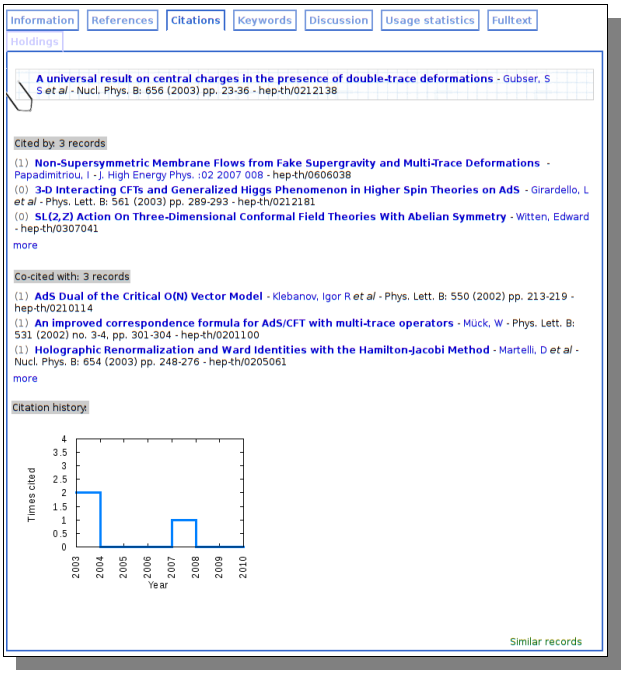
Searching and browsing records

A powerful search engine resides at the core of Invenio. Supporting a Google-like syntax, it can be used to dig into the repositories in a very easy way, while allowing the more experienced users to take advantage of many advanced features such as regular expressions, range queries, selection of meta-data, and many more. Results can be clustered by collections and ranked with respect to word similarity, number of citations, downloads based rankings, and any other pluggable algorithm.

Any search query can automatically turned into an RSS feed to add to a personal news aggregator.

Thanks to integrated support for MODS and UnAPI standards, search results can be stored in Zotero, a powerful Firefox plug-in useful to organize a personal digital library.

The system offers users powerful citation and author analyses (e.g. co-authorships, most used keywords, etc.).



Collaborative tools

Invenio integrates several collaborative and typical web 2.0 tools such as baskets, to store and share records or external resources with other users, personal email notification alerts, record commenting and reviewing, internal web-based messaging tools, approval/reviewing work-flows.

Exporting records

Data and meta-data can be exported in several ways, the most important being through OAI-PHM protocol and by means of a pluggable exporting module (used today to generating sitemaps and exporting to Google Scholar).

Authentication and authorization

While offering for free local accounts management, Invenio can be integrated with institutional user databases to delegate the authentication phase and to exploit externally available user details.

Using the role based access control (RBAC) model of Invenio one can define a list of roles, attach users to particular roles, and grant rights to perform actions such as view-restricted-collections based on user group membership, user IP address, and more.

Circulation

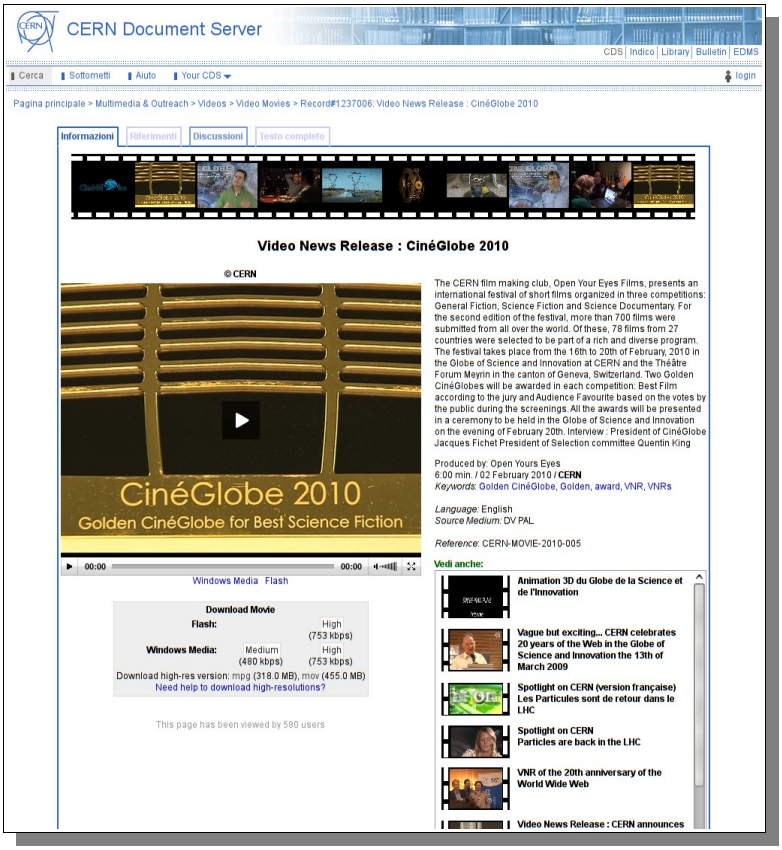
Invenio supports the operation of a physical library by integrating a circulation module, as well as by supporting inter library loans. Users can submit loan requests, and follow the request status, the librarians have a web interface to accept/reject loan requests, to send overdue letters, to set up sub-libraries, etc.

Creating Journals

As an example of an application that has been built on top of the core Invenio, the web journal module lets repository administrators create online web journals with regular issues built from articles stored in the repository itself. A specific appearance can be set up in order to have a layout providing journal-like reading experience. This is currently used today for producing the official CERN Bulletin, which is issued every two weeks.

Conclusions

Invenio is a mature and stable digital library software with more than 7 years of open source history. It is targeted towards mid-to-large software repositories (~2M records) containing diverse material (papers, books, photos, videos). It has attracted interest of especially large subject-based repositories or large library networks. It is designed to cope with diverse requirements and to provide high performance.



The Invenio Development Team