Nové funkce a technologie v současných a budoucích verzích Invenia

Kunčar, Jiří
2012

New Features and Technologies in Current and Future Invenio Versions

Jiří Kunčar and Tibor Šimko, for Invenio collaboration
{jiri.kuncar@cern.ch, tibor.simko}@cern.ch

October 17, 2012

Abstract

The aim of this paper is to introduce new features and technologies in upcoming versions of the Invenio software suite. Invenio was originally developed at CERN (European Organization for Nuclear Research) and covers all aspects of digital library management or document repository on the web. We focus on describing improvements of full-text search using external ranking tools, multimedia management, circulation and holdings module, and new user interface built on top of new technology stack using a Python micro-framework, modern templating engine and powerful SQL toolkit.

1 Introduction

CERN The European Organization for Nuclear Research in Geneva is the world's largest particle physics laboratory established in 1954 by 12 European countries. CERN community now includes 20 European member states with over 40 additional participating observers (states and organizations) with about 10,000 visiting scientists from 608 universities and 113 nationalities coming to CERN to work on their research every year [1].

Throughout almost 60-year long history, CERN has established a solid reputation in scientific research and the work at the laboratory is pushing the boundaries of particle physics and related industry. The main current research program at CERN is to understand the basic constituents of matter using the *Large Hadron Collider* (LHC), the world's largest and most powerful accelerator, smashing particles together in a ring of superconducting magnets with a circumference of 27 kilometers located 100 meters under the ground.

Perhaps the most important invention that has been driving our lives over last decade is the *World Wide Web* (WWW) invented by Tim Berners-Lee in the late 1980s and early 1990s to improve communication and collaboration services for the world-wide particle physics community. The Web has opened doors for many innovative applications, keeping the standards open for all.

Invenio is an open source web-based application that implements an integrated *digital library system* or *document server* [2], and is used at CERN to run the CERN Document Server Institutional Repository (CDS) – currently one of the largest institutional repositories worldwide with more than 1 million of records. More than 10 years of development has proven stability and maturity of the software used by approximately 50 institutions (CDS, INSPIRE, ILO, NTK, EPFL, etc.). Invenio is free software project with *GNU General Public Licence*. It is based on *Linux, Apache/WSGI, Python*, and *MySQL*.

Open standards such as MARCXML and OAI-PMH 2.0 have been adopted in the very beginning to ensure interoperability with other digital libraries. The Invenio and CDS teams cooperate with INSPIRE, the next-generation High Energy Physics information system, that has been built in a world-wide collaboration among *CERN*, *DESY*, *Fermilab*, and *SLAC*, and interacts closely with *arXiv* and *NASA-ADS*. The modular architecture and high adaptability enables Invenio to serve a wide variety of requirements, from a multimedia digital object repository, through a web journal, to a fully functional mid-to-large scale digital libraries and repositories with 100K-10M records.

Last year was very fruitful for Invenio community. The major version of Invenio 1.0.0 has been released as well as several bugfix updates to the 0.99 series. We were organizing the *1st Invenio User Group Workshop* where lots of new ideas and improvements were discussed. Recently, we have introduced new minor release 1.1 (presented in Section 2) with the most demanded features. Furthermore, the underlying technology of Invenio has been strengthened by adopting new web and UI toolkits (presented in Section 4).

2 Invenio 1.1

Invenio 1.1 is a new release bringing many features over 1.0 release series. There were more than 1200 commits in last 2 years work of development. We shall list here some of the most notable novelties of Invenio 1.1.

Invenio upgrader enables automated deployment of new Invenio releases. An administrator of an Invenio instance can run this command to see which changes will be applied by an upgrade and whether the upgrade can run unattended or require adaptations by a human. Before each upgrade of an existing installation, the dependency graph of individual upgrades is calculated and checked. The upgrade engine also supports having several independent graphs; you normally want one graph for Invenio and one for your local overlay repository.

The upgrade engine will run upgrades in topological order (i.e upgrades will be run respecting the dependency graph). The engine will detect cycles in the graph and will refuse to run any upgrades until the cycles have been broken. An important feature of the upgrade engine is the ability to run upgrade pre-checks prior to actually being installed into site-packages. This allows a system administrator to validate the upgrades prior to running the installation command, where it would be too late to discover issues and roll-back.

Circulation and holdings module enables you to manage the circulation of books (and other items) in a traditional library. It is a result of an effort in lowering infrastructure maintenance costs by grouping several library services on a single server and provides an alternative to a commercial system reusing the existing technologies and concepts of Invenio.

Invenio circulation and holdings system is used at CERN by about 10k users (borrowers) that can borrow a wide range of 35k books. The librarians with correspondent rights to Invenio system can manage acquisition of institutional items (anything that can have a barcode), loans (including ILL), and user requests.

Multimedia encoder is a new multi-purpose video treatment module allowing the execution of video transcoding, frame extraction, and metadata handling as scheduled tasks. The core functionality is covered by FFMPEG wrapper. There is also new interface for video player and reworked page layout coming with the module.

Author identifier is new author disambiguation and claiming module developed within INSPIRE use case. The algorithm that tries to find matching between virtual authors (the author as it appears on a document) and real author (ideally a real individual researcher) is based on textual similarity, affiliation history, topics of papers, etc. The module also allows individual researchers to claim ownership of their papers via web interface.

Sorting buckets are used to implement more efficient sorting algorithm for large repositories. To achieve its goal it creates several sorting buckets that hold record identifiers containing approximately the same number of items in memory. When a sorting method is selected, search results are consecutively intersected with individual buckets until requested number of items is found. This approach will allow user to define multiple sorting options without unnecessary performance penalties, because only limited amount of items has to be really sorted.

3 Invenio Master Branch

Many new features get into the *master* branch after passing all necessary tests and reaching desired code quality. The *master* branch serves as a stabilisation platform for these features. Stabilized features should appear in Invenio 1.2 release series.

Record field model is new flexible logical field and data model description system that has been introduced in order to permit non-MARC master formats (e.g. EAD) in Invenio. It is adding support for virtual fields and provides an easy API to access this information. Finally there is centralized definition and documentation for all the fields of the record data model of an Invenio instance.

Improved search for restricted records now automatically includes restricted results into user searches depending on rights of authenticated user. For example if you are logged in and you are searching from the main page, you will get records from collections to which you have appropriate rights without having to explicitly pre-select those concrete restricted collections to searches.

External ranking generic bridge for metadata indexing and word similarity ranking tool has been implemented[5] . The work was based on the preliminary successful fulltext search integration of *Solr* and permits to implement enhanced ranking techniques and improves scalability of the word similarity ranking method. The code has been stabilized and tested on CERN Document Server production data.

4 Invenio Next Branch

Invenio has reached the point where it is useful to rewrite part of the code base using new software stack in order to keep flexibility, manageability of growing number of modules, and speed-up prototyping and development of new ones. Moreover it is desirable to ensure independence from the database system. Hereafter, we describe technologies and their function in the project used in the *next* branch.

4.1 Adopted Technologies

SQLAlchemy is one of the most used Python SQL toolkits with an optional *Object Relation Mapper* (ORM) component that provides the data mapper pattern, where classes can be mapped to the databases in open ended, multiple ways. It allows the object model and database schema to develop in a cleanly decoupled way from the beginning [3]. SQLAlchemy allows to express wide range of SQL statements; any of these units can be later composed into larger structures.

The object models were designed in a way permitting dynamic properties with strong checks ensuring data consistency. For example, the *User* class does not only represent the user table, but also allows to see which messages user has received, etc.

Jinja2 is a full-featured template engine with an expressive language that forces the strict separation between business logic and presentation[7]. It allows creation of reusable building blocks across the whole system that reduces the need for repeating code. During rewriting *user interface* (UI) templates, we have focused also on modern responsive design for different devices and screen resolutions using CSS front-end framework.

Twitter Bootstrap is an open-source collection of a series of LESS[9] style-sheets that provides style definitions for various UI components[4]. In addition to the standard HTML elements, Bootstrap contains other widely used interface elements and JavaScript based plugins to extend the functionality of existing interface elements.

WTForms is a Python library to automatize process of form definition, generation, and validation. Field generation can be easily customized using *widget* templates.

Flask is a powerful lightweight microframework for web applications based on Web Server Gateway Interface (WSGI) toolkit called *Werkzeug*[6, 8]. It uses a concept of *blueprints* for factorizing an application into a set of components that can be registered on an application at a URL prefix and/or subdomain even multiple times. Blueprints are preferred method used by Invenio developers to implement local custom modules allowing a clean separation of custom overlay from Invenio core. Flask is also integrated with *Jinja2* templating system and with *SQLAlchemy* (via an extension).

Redis is an open-source in-memory key-value data store with an optional data persistence and master-slave replication support. The type of data stored in Redis is not limited to strings, but several other abstract data types are supported – lists, sets, sorted sets and hashes. The basic functionality is available through general *Flask-Cache* extension and will be required for *facets* or *user sessions.*

4.2 Selected Features

Search interface has got auto-completion feature for search fields and possibility to use indexes or knowledge bases to help user with correct spelling of author, journals, etc. Moreover new interactive forms for building search queries have been developed as a replacement for the advanced search interface. This new forms allow even a novice user to construct complex queries without having to know the peculiarities of the advanced search syntax.

Faceted search is technique for refining search results using information organized according to classification criteria. It enables users to narrow number of choices in each classified criteria dimension of information space after search. The results are combining initial text search query with selected criteria. Users still have a freedom of defining their own advanced searches; moreover they can get an inspiration from generated queries while they are using facets.

Bundling static files support has been added to *Jinja2* templates to improve page load time. Many modules has their own JavaScript (JS) libraries or additional cascading style sheets (CSS) that have to be loaded for certain pages. Therefore browsers have to download multiple files and for each of these files, an HTTP request is sent to the server. This HTTP overhead causes delay before the page can be rendered. The goal of this extension is to allow developers to keep their JavaScript and CSS files separated to make the development easier while automatically optimising JS and CSS bundles and loading for production conditions.

5 Conclusion

Since its first release in 2002, Invenio user and developer community has been growing steadily, attracting mid-to-large document repositories both inside and outside the high-energy physics domain. Nowadays there are over 30 installations world-wide totalling over 4 million of documents. The developer community spread outside of Europe and reached more than 35 core developers, temporary code contributors, and voluntary interface translators in 2011.

In this paper we have presented some of the new features coming with Invenio 1.1 release series that was introduced in 2012. We have also described the current major effort at strengthening the underlying technology code base and at modernising the user experience. The current efforts will permit easier customisation of the core behaviour of the software and its smoother adaptation to diverse specific needs of the growing user and developer communities.

References

[1] CERN (Where the web was born). `http://public.web.cern.ch/public/en/About/Web-en.html`, 2008. [Online; accessed 7-October-2012].

[2] Invenio. `http://invenio-software.org/`, 2012. [Online; accessed 8-October-2012].

[3] SQLAlchemy (The Database Toolkit for Python). `http://www.sqlalchemy.org`, 2012. [Online; accessed 7-October-2012].

[4] Twitter Bootstrap. `http://twitter.github.com/bootstrap/`, 2012. [Online; accessed 8-October-2012].

[5] P. O. Glauner, T. Simko, and H. Vogelsang. *Enhancing Invenio Digital Library With An External Relevance Ranking Engine. oai:cds.cern.ch:1456329.* PhD thesis, Karlsruhe U., 2012. Presented 2012.

[6] A. Ronacher. Flask (A Python Microframework). `http://flask.pocoo.org/`, 2012. [Online; accessed 8-October-2012].

[7] A. Ronacher. Jinja2 (The Python Template Engine). `http://jinja.pocoo.org/`, 2012. [Online; accessed 7-October-2012].

[8] A. Ronacher. Werkzeug (The Python WSGI Utility Library). `http://werkzeug.pocoo.org/`, 2012. [Online; accessed 15-October-2012].

[9] A. Sellier. LESS (The dynamic stylesheet language). `http://lesscss.org/`, 2012. [Online; accessed 15-October-2012].