INVENIO INTEREST GROUP PROPOSAL - *OPEN REPOSITORIES 2014*

INSPIRE: contributions to Invenio from the leading High Energy Physics (HEP) platform

Javier Martin Montull <javier.martin.montull@cern.ch> - CERN

Keywords: information system; digital library; open source; aggregation;

INSPIRE (http://inspirehep.net/) is an information system for the global High Energy Physics (HEP) community, supported by a collaboration between four major particle physics labs: CERN, DESY, Fermilab and SLAC. INSPIRE is used regularly by around 50,000 HEP scientists worldwide, in order to search for publications, references, authors, institutions, conferences, job offers, and more.

INSPIRE currently holds over 1 million bibliographic records. Part of this content is inherited from its predecessor SPIRES which was the first website available in the US and ran for over 40 years. With support from the collaboration, all content from the SPIRES database was imported and its ageing infrastructure replaced by the digital library software Invenio (http://invenio-software.org/). The use of Invenio provided faster out of the box results, a variety of search and display options, searchable fulltext, better harvesting capabilities and much more. The INSPIRE website was then launched in 2008.

INSPIRE aggregates information from different sources. A daily harvest is done from the partner repository arXiv (http://arxiv.org/) via OAI-PMH protocol; several publisher feeds are also ingested periodically; and users can propose new content to be added to the system or suggest corrections.

Once the content has been ingested, automatic tools that are part of Invenio try to match the new records against the database content in order to avoid duplicates, then automatic reference extraction is performed and other jobs are run in order to enrich the metadata of the new records and extract useful information about the authors.

After all the automatic processing has taken place, a distributed team of catalogers (from the four participating laboratories) perform manual curation using Invenio's cataloguing tools. These tools have been improved to meet the INSPIRE needs, while allowing the entire Invenio community to benefit from it.

The proposed presentation will provide an overview of all the functionalities that INSPIRE has contributed to the Invenio software.

The first of these is the author disambiguation functionality and automatic author profiles. INSPIRE uses machine-learning algorithms in an attempt to automatically aggregate papers from the same person under a user profile page. This is extremely challenging due to the variation in a single author's signature, e.g John Doe; Doe, John; Doe, J. As this process cannot be 100% accurate based only on machine algorithms and metadata, a crowd-sourced "paper claiming" feature was also developed and is now part of Invenio. Since

research output is often used as an indicator of academic impact, this type of information is extremely valuable for the HEP community.

Also of great importance is accurately keeping track of references. Within INSPIRE, the automatic reference extraction from PDF documents has been improved and integrated into the Invenio software, thus enabling the creation of a citation graph that would span almost the entire scope of HEP literature.

Other interesting developments are related to the back-office cataloging tools that INSPIRE's staff use to clean the metadata and handle user requests. These include a record editor, a batch record editor, a record merger and a programmable metadata cleaner. All these developments have been integrated in Invenio over the past few years.

Finally, the presentation will outline future development plans for INSPIRE, focusing on migration to the next version of Invenio. This includes porting modules to the Flask/Jinja2/Bootstrap technologies, and making use of the new BibField record representation in Invenio.