CDS Invenio - a software solution for National Repository of Grey Literature

Tomáš Müller
National Technical Library,
Prague, Czech Republic
tomas.muller@techlib.cz

Third Seminar on Providing Access to Grey Literature
December 8, 2010

Abstract:

For retrieving, preserving and managing digital documents of gray literature and its metadata a sophisticated software system must be used. CDS Invenio is a software solution for the needs of NRGL – for the central repository, which collects and the grey literature and makes it accessible and for the co-working organizations, which produce the grey literature. Its complexity, flexibility and open-source solution ensure fulfilling of nearly all requirements that NRGL have for this system.

Contribution:

The grey literature

There are many definitions of what the grey literature is. One of the most famous definitions appeared in 1997 in Luxemburg and it was expanded in New York in 2004. It says that the grey literature is: "information produced on all levels of government, academics, business and industry in electronic and print formats not controlled by commercial publishing i.e. where publishing is not the primary activity of the producing body"[1], so under the term “grey literature” we can imagine things like reports (i.e. annual reports, research reports, final reports from projects, …), conference materials (i.e. posters, presentations, papers, …), theses (i.e. bachelor theses, master theses, rigorous theses, …), even common correspondence, like letters, e-mails, web blog posts and so on. Some of these documents might contain some very valuable pieces of information and that is the reason why this phenomenon is so discussed nowadays.

The NRGL project

The good news is that grey literature probably contains a lot of useful information. The bad news is that finding some particular grey literature document we are interested in is very difficult. These documents are usually scattered in local repositories and there is no way how to search among all of

[1] Luxembourg, 1997 - Expanded in New York, 2004, dostupné na WWW <http://www.greynet.org/index.html>

them at the same time and that is why NRGL was created. [2]

The main purpose of NRGL is to create a central national repository of grey literature that will collect metadata and full texts from local repositories and ease the access to actual documents. When all these data are in one place it's much easier to find what we want.

NRGL tries to find organizations that produce potentially valuable grey literature (such as colleges, science institutes, libraries …) and that are interested in sharing it. Metadata usually contain some sensitive information so it is necessary to make a contract with the institution. Then the institution may insert their records directly into NRGL or the metadata will be harvested from their local repository into NRGL.

CDS Invenio

System overview

CDS Invenio is document management system developed by Swiss company CERN. It is free and open source solution for library systems and repositories.

The architecture of the system is modular, most parts are written in Python programming language, so it is not difficult to extend it if necessary.[3]

Performance of the system is well optimized. All the most demanding outputs are cached to minimize the communication with database, what increases the speed of the system. Invenio also uses various indexes. Once created, searching is very fast. All these features are contributing to user friendliness but on the other hand they make the administration of the system much more difficult. Practically it is not possible to run this system without the knowledge of Python or al least some remotely similar language. It is also very important to keep in mind that most changes in the configuration won't apply immediately due to caching. For instant result it is necessary to refresh the cache manually. The biggest down of this system is that the current version is still 0.99.1, so some parts of the system is still being developed and the documentation is rather brief and some parts are outdated. A stable version 0.99.3 and development version 1.0 RC 0 were released recently, but there are a lot of changes in our system, so we are planning to wait for stable 1.0 version to import all modifications there.

The administration occurs both via web browser and via command line of the computer on which Invenio runs. Web browser administration usually allows the administrator to set up the system, while command line administration is used mostly to run various tasks.

---

[2] PEJŠOVÁ, Petra. Národní úložiště šedé literatury (NUŠL). Čtenář : měsíčník pro knihovny. 2010, roč. 62, č. 5, s. 176-180. Available from WWW: <http://ctenar.svkkl.cz/clanky/2010-roc-62/05-2010.htm>.

[3] *CDS Invenio* [online]. 2010 [cit. 2010-12-05]. About Invenio. Available from WWW: <http://invenio-software.org/>.

Figure 1: User web interface of Invenio in NRGL

Invenio in NRGL is installed on virtual machine, which can be exported and provided to organizations that want their own repository and don't have one. Using exported virtual machine will save them from installing a configuring Invenio, what is hard and tedious job. Virtual environment is currently provided by VirtualBox.[4]

Data structure

The records are internally stored in MARC 21 format. Each record consists of several fields defined by 3 place number and 2 indicators. Each indicator is a single number or blank. Every field has one or more subfields, which are defined by single character or number. The meaning of the fields and subfields are charted.

The records in NRGL divide in two ways - by document type and by institution. A division unit is called collection. The document types are following[5]:

[4] *VirtualBox* [online]. 2010 [cit. 2010-12-05]. About VirtualBox. Dostupné z WWW: <http://www.virtualbox.org/>.

[5] *Typologie dokumentu NUŠL* [online]. 2010 [cit. 2010-12-05]. Národní úložiště šedé literatiry. Dostupné z WWW: <http://nusl.techlib.cz/index.php/Typologie_dokumentu>.

- Theses
  - Habilitation theses
  - PhD theses
  - Rigorous theses
  - Master theses
  - Bachelor theses
- Reports
  - Survey reports
  - Grant reports
  - Final report of the project
  - Interim report of the project
  - Statistical reports
  - Technical reports
  - Research reports
  - Annual reports
- Copyrighted Writings
  - Preprints
  - Papers
- Trade Literature
  - Product Catalogues
  - Guides
- Conference Materials
  - Posters
  - Presentations
  - Proceedings
  - Programs
  - Articles
- Study Materials
  - Course Synopses
  - Exam Questions
  - Teaching Transcripts

Assigning some record to a collection is indirect by logical field. Logical field is defined in BibIndex module. We select a field(s) and/or subfield(s) which we want to index and some name of the index. For example we want to create a logical field called ‘collection’ for 980__a (field 980, both indicators blank, subfield ‘a’). From now on we can search among the records by values in field 980__a. Now we may for example set that collection ‘Preprints’ is defined by ‘collection’ logical field and by value ‘preprints’ (correct syntax is ‘collection:preprints’). Now all records which have in field 980__a value ‘preprints’ (case insensitive, ignores accentuation) will appear in ‘Preprints’ collection. Like that we can select whatever filed to search by or to assign to collection by.

Data acquisition

We can deliver metadata and full texts into CDS Invenio in three ways – submit using web form, submit using e-mail and harvest from another repository. In NRGL only web form submission and harvesting is used.

Submitting through web form consists of two parts. First – creating a form(s) (there can be more pages of the form) and second – construct a

sequence of functions and their arguments that will process the data obtained through the form. First we have to define all the elements we want to use in a form. In fact the elements are classic form objects like text input, select box etc. There is of course more set up like element name, description and so on. Then we have to arrange these elements into a form with some labels, setting whether the element is optional or mandatory and so on. Now we have a complete form which we can fill in, so we have to create a sequence of functions which will do things like creating system number, renaming submitted files and moving them to storage, creating the actual record, upload record etc. That should result into record in MARCXML format, which we can upload. We can even write our own functions. We can define several document types and their subtypes and for each type we can have a separate form while the elements are still the same. Apart from submitting a new record, these forms can be used for example for editing some record or just adding a file to existing record and much more.

Harvesting records from another repository is carried out by OAI-PMH protocol.[6] It sends record in some XML format (mostly in DC or MARCXML) through HTTP protocol by batches of usually 100 – 1000 records. These records must be converted into MARCXML format before they can be uploaded to the system. The most convenient conversion is by XSLT. Converting large number (e.g. thousands) of records may be difficult sometimes, particularly when the harvested data are not consistent. The most important is to create all tags which are used to assigning records to collection exactly right, which can be really hard sometimes.

CDS Invenio can also play the role of data provider. We can specify a set of data which will be exposed for harvesting to some other repository using once again OAI-PMH protocol. NRGL will use this feature for joining the international project concerning grey literature such as Open Grey.

Other utilities

There are a great number of features offered by CDS Invenio, so there will be mention only those that are widely used by NRGL.

Security in CDS Invenio is solved by classic role-based model. Basic element in security module is action with its parameters. Actions group into roles. Roles are assigned to users directly to user accounts or indirectly by "firewall-like" settings. Firewall-like role assignment is very powerful feature which allows us to set a role to user e.g. by user's IP address. This allows us for example to restrict access to some collections or full texts to only certain networks.

BibRank module is capable of computing some "special" indexes like citation index or word similarity. This enables the option of searching similar records, which can be very useful feature e.g. for research jobs.

There is a need of running many periodical tasks in Invenio, such as indexing, harvesting, cache refresh, cleanup tasks and many more. Invenio runs its own task scheduler (BibSched module), which is used for every task in the system. So if someone submits a new record, the uploading is put in a queue in a scheduler, it is not processed immediately. Many changes in Invenio will

[6] *The Open Archives* [online]. 2008 [cit. 2010-12-06]. The Open Archives Initiative Protocol for Metadata Harvesting. Dostupné z WWW: <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

appear when it's their turn in the scheduler, not right away, which may seem a little odd to the users, but it helps to maintain optimal performance.

Enhancements

Even so complex system like Invenio can't offer everything. There are some things that must have been made-to-measure.

When converting harvested data we can't know the system number of the records in advance. This number is assigned when the record is being uploaded. Yet we need it for creating a record identifier. This problem was solved as post-process task. This task finds all records without identifier and fills it in.

Similar problem is with date of uploading of the record and date of modifying of the record. These two entries are stored in the database but we need them in MARC as well. So similar task as mentioned above was created, that will keep these two data field up to date.

Another missing feature is the possibility of exporting and importing records in the system, so this feature had been created as well. Export script dumps that part of database that contains data of the records and copies all the full text in some backup directory. Export script works exactly oppositely. It takes dumped part of database and loads it up and copies full texts back into its place.

Currently and automatic indexation tool is being developed in Invenio. It shall be analyzed the full text and based on its content it will suggest some keywords to us. This feature should help a lot with document description.

Conclusion

So far we have a fully operational digital repository with 3 harvested repositories – Academy of Science, University of Economics and our institutional repository (altogether about 42000 records) and about 10 manually inserted records. We created a manual for Invenio installation, collections management and WebSubmit templates a we are working on FAQ, where we want to describe some important tasks which are hard to understand from the official documentation.