Seminar on Providing Access to Grey Literature 2011: 4th year of the seminar focused on storage and making the grey literature accessible, 25. 10. 2011 [online]. Praha: National Technical Library, 2011. Available at WWW: http://nusl.techlib.cz/Sborniky. ISSN 1803-6015.

CDS INVENIO FOR NATIONAL REPOSITORY OF GREY LITERATURE – MIGRATION TO VERSION 1.0rc0

TOMÁŠ MÜLLER

tomasuv.mail@gmail.com

National Technical Library, Czech Republic

Abstract

Until the beginning of the year 2011, the National Repository of Grey Literature used the system CDS Invenio, version 0.99.1 as its digital repository. At the end of the year 2010, CERN released CDS Invenio, version 1.0 release candidate 0. This version has so many useful features for NRGL, so we decided to take that brave step and introduce this still testing version into the main run.

Keywords

CDS Invenio, Software, Software development, Data migration

Initial state

The National repository of Grey Literature (NRGL) has been using system CDS Invenio[1] for their digital repository since September 2009. The choice has been made based on the list of requirements and public tender made in 2008 and 2009.

The latest version of system CDS Invenio back then was 0.99.1. It was installed on SUN (physical) server with Solaris[2], which was the only one of its kind in National Technical Library (which runs the NRGL project) and that induced some difficulties with backup and general technical support. There was a virtual environment VirtualBox[3] installed on physical server SUN, where a virtual machine ran. On this virtual machine was installed the system CDS Invenio 0.99.1 under the OS Linux Debian[4]. This virtual environment was chosen because it allows us to export the virtual machine a distribute it to our co-working institutions but generally, it is more designated for desktop application than for servers.

The version 0.99.1 is still a pre-release, so there is likely a number of known issues.

Some modules were not fully completed - the edit record interface used temporary files which sometimes caused errors and all new fields were added twice, the upload script required the records in MARCXML unnecessarily precise (e.g. no blank fields) otherwise it raised exception, there was no comfortable way how to manage digital documents and so on.

A big problem in this version was documentation, which was rather brief, outdated or it wasn't present at all. Many mechanisms had to be deduced from source code (in brighter cases from source code comments) and by method trial-and-error.

[1] Invenio [online]. [cit. 2011-11-25]. Available at WWW: <http://invenio-software.org/>.

[2] Oracle [online]. [cit. 2011-11-25]. Solaris Overview. Available at WWW: <http://www.oracle.com/us/products/servers-storage/solaris/overview/index.html>.

[3] VirtualBox [online]. [cit. 2011-11-25]. Available at WWW: <https://www.virtualbox.org/>.

[4] Debian [online]. 2011-11-25 [cit. 2011-11-25]. Available at WWW: <http://www.debian.org/>.

In rare cases happened, that the system froze completely and had to be restarted, although it was hard to say whether it was failure of Invenio, operation system (on virtual machine) or the virtual environment.

The installation is quite complex and there are missing some important information. Anyone who wants to install it needs quite good knowledge of the OS and its components (Apache, self-signed certificates, etc.).

Before the transition to higher version there were about 20000 record and less than 50 digital documents.

New version

The version 1.0 was promised since autumn 2010, but only version 0.99.3 was released then. We didn't accept this version because it seemed that it didn't bring anything important for us. At the turn of the year 2010/2011 the version 1.0rc0 was released. RC stands for release-candidate. That means, that all features which will be in 1.0 release are implemented and if no major bugs are found, it will become the first official release.

We tested the version 1.0rc0 with very promising results and because the progress to release of the version 1.0 seemed to be very slow, we decided to accept version 1.0rc0 and install it under standard virtualization platform used in our institution (Xen[5]).

Benefits of the new version

The general run of the system is much smoother. The exceptional states are very rare and so far the system never froze completely. The record editor is brand new, it uses modern technologies like JQuery[6] and JQuery UI, so the work with record editor is very interactive, smooth and user-friendly. Completely new feature is Document File Manager, which allows us to manage digital documents associated with individual records via the Internet browser in very user-friendly way. The indexes were also improved. Previously only the words and sentences were indexed, now the word pairs are indexed as well. There are also several new modules like BibJournal (the web journal), BibCirculation (management of the physical copies), Multi-record Editor, BibMerge (record merging) and some others.

Migration

There had been made a lot of changes in Invenio (in source code, configurations, …) so rather then just upgrade current installation it was better to migrate the data to new freshly installed and configured system. Besides we wanted to switch to another virtualization platform.

The migration can be divided into 4 stages:

Metadata
Digital documents
Configurations
Modified source code

Metadata

Migration of the metadata should have been the easy part, but in the end it caused most troubles. The main point here is that record must keep their system number (ID). In the previous version

[5] Xen [online]. 2011 [cit. 2011-11-25]. Available at WWW: <http://xen.org/>.

[6] JQuery [online]. 2010 [cit. 2011-11-25]. Available at WWW: <http://jquery.com/>.

there were several soft-deleted records (that means they were just marked as deleted, everything behave like they were deleted, but in fact they are still physically present), so we need to migrate them as well to prevent re-numbering.

Invenio in default doesn't offer any known tool to extract record from database “as it is” in all cases (deleted records), so the easiest way of migration is to import that part of database where the metadata are stored. Later it had been shown, that this was a mistake because it caused some unpredicted behavior, specifically in exposing OAI set – it took about 2 or 3 second to every third or fourth record to format, other records were formatted thousand times faster. So in the batch of 500 records the time delay was about 8 minutes (instead of a few seconds). The true nature of this problem had never been discovered but after second try with other method of migration this problem didn't appear.

On the second try we decided to follow Invenio mechanisms and upload the records in standard way. For this a special function must have been written, that can just extract the record from the database and prepare it for re-upload.

Digital documents

The digital documents are kept separately from the metadata (metadata are in the database and the digital documents are on the file system), so we need to carefully migrate the files and preserve consistency at all cost.

Using Invenio API for working with digital documents would be too complex (and prone to mistakes) and going around and using the database access was rejected due to previous experience. We used the fact that both repositories were operational and the matadata in consistent state, so we could just harvest the full texts from the old repository to the new one.

Configurations

Configurations (like format templates, format elements, conversion sheets) in most cases could be used as it was in the previous version, some of them needed some light modifications.

Modified source code

Modified source code can't be used as whole files, because the rest of the file most likely changed with the new version. Only the code snippets with the actual change can be used after some modification if needed. All these changes must have been done manually, because the nature of this part of migration is so complex, that it can't by done in batch and safe.

Further development

The system CDS Invenio is constantly being developed in NTK (besides regular development in CERN). Small but important changes in the source code are made to improve the compliance with the needs of the National Repository of Grey Literature.

Advanced conversion with Python

Invenio allows us to convert the records in two ways – XSLT[7] and BFT.

BFT (BibFormatTemplate) is an Invenio tool for transforming text input. Its primary purpose is to transform user input (from module WebSubmit[8]) to MARCXML format, but it had been generalized

[7] XSL Transformations (XSLT) [online]. 1999 [cit. 2011-11-25]. Available at WWW: <http://www.w3.org/TR/xslt>.

[8] WebSubmit module provides the tools for creating interactive interface for creating record and inserting them into

for general usage. The power of this tool is very limited and syntax of the templates is quite complex. It is not very suitable for general purposes.

XSLT (Extensible Stylesheet Language Transformations) is a tool designated for transformation data in XML format to another format (most often into another XML or HTML). It is quite easy to use but more complex problems are very hard to express.

Because there was no really powerful tool for solving complex problems, a third option was enabled – pure python script. The expressive power and number of libraries to use greatly outweighs the disadvantage of greater complexity of the programming language itself. The records can be converted in very complex way if necessary.

This new feature was implemented even in formatting module and for processing user input from WebSubmit module.

Automatic indexing with PSH[9]

The automatic indexing service was implemented in NTK. It is based on Maui indexer[10], the indexing model for Czech language was created using about 120 manually indexed records with digital document. There are over 13000 subjects in PSH so 120 articles is still pretty low. The indexing model will be further improved.

Automatic indexing process is now a part of WebSubmit. When user submits his paper, the automatic indexation suggests the user some subjects and he can then add some more or delete them.

Additionally there is an standalone web interface available at http://invenio.nusl.cz/indexer/ where can anyone submit his own text and receive automatically assigned PSH subjects.

Local installation of CDS Invenio

If some cooperating institution want to have its own digital repository, we offer them pre-configured CDS Invenio to install. Formerly we used exported virtual machine from VirtualBox with the whole operation system. The cooperating institution needed just to install VirtualBox on whatever operation system they had.

This solution proved to be very unreliable, because there were very serious problems with importing the virtual machine and the cooperating institutions often didn't want to install VirtualBox (e.g. they were using some other virtualization platform).

We decided to create an installation script, that will handle all the installation routine and some basic settings. The cooperating institutions must provide just a computer with OS Debian or Ubuntu (or similar) installed (no matter whether physical or virtual machine) and the installation script will navigate the administrator through the installation process.

Troubleshooting

The 1.0rc0 version of Invenio wasn't without serious errors. Fortunately all of them were quite easy to fix.

The most problematic part was OAI interface[11]. Even if there was set "persistent" deletion strategy

CDS Invenio system

[9] *National technical library* [online]. 2011-08-02 [cit. 2011-11-25]. Polytematický strukturovaný heslář. Available at WWW: <http://www.techlib.cz/cs/14-psh>.

[10] Maui-indexer : Multi-purpose automatic topic indexing [online]. [cit. 2011-11-25]. Available at WWW: <http://code.google.com/p/maui-indexer/>.

[11] The Open Archives Initiative Protocol for Metadata Harvesting [online]. 2008-12-07 [cit. 2011-11-25]. Available at

(record stays in OAI set, but it is marked as deleted), the deleted records disappeared from the set. This was caused by using the same search function for retrieving records to OAI set as for classical searching. This search function by default ignores "deleted" records. Instead of this function a SQL query was used for this purpose.

Another problem with OAI was with function that organized records to OAI sets. The whole philosophy of this function was defective, so it had to be re-engineered.

Another big problem was full text versioning. Whenever a new full text come to a record, Invenio keeps the previous versions of this document. The problem is that Invenio don't check whether it already have it or not, so anytime a "new version" of full text came (even if it is exactly the same as some of the previous versions), it was added as new version. And it may happen quite often. This was solved by checking the MD5 checksums[12] of all present digital documents and comparing it with MD5 of the incoming file and add it only if the MD5s don't match.

The last serious problem solved was about search options. It was more an unpleasant implementation than error. There are several collections in Invenio (in case of NRGL over a hundred) and you can select a set of search fields for each collection. The problem is that you have to select it for each collection separately. There was no way how to do it globally. When no search fields were selected for actual collection, some set of default search fields was used (but it may not correspond to our settings of indexes). This problem was solved by making up a rule, that if some collection have the search fields unset, it accepts the settings from parent collection. In this way it's enough to set only 1 (root) collection and the settings will spread to all subordinate collections.

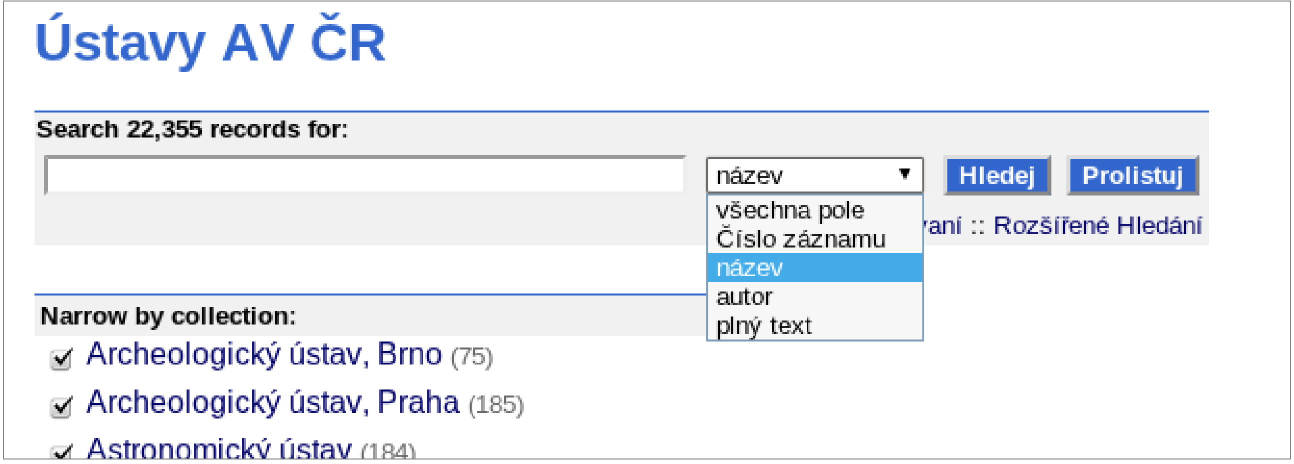

Fig. 1 Search fields in Invenio

Conclusion

All the data, settings and modifications were successfully transferred from the old repository to the new one.

So far there is over 67.000 records and over 200 of them with attached digital document. The development of the repository and spreading of our cooperation network is still in progress.

WWW: <http://www.openarchives.org/OAI/openarchivesprotocol.html>.

[12] The MD5 Message-Digest Algorithm [online]. 2011 [cit. 2011-11-25]. Available at WWW: <http://www.fastsum.com/rfc1321.php>.