NTK

50°6'14.083"N, 14°23'26.365"E
Národní technická knihovna
National Technical Library

Cestovní zpráva

# Invenio User Group Workshop 2012
# CERN, Ženeva, Švýcarsko
# 7. 5. 2012 – 9. 5. 2012

Účastnice:
Mgr. Jana Slouková (jana.sloukova@techlib.cz, DNTK 51)

Květen 2012

## Účel cesty

Cílem zahraniční služební cesty byl švýcarský CERN, kde je mimo jiné vyvíjen open source systém pro správu digitálních dokumentů Invenio. Tento software původně vzniknul pro potřebu tamní knihovny, postupem času se však rozšířil i do jiných institucí (včetně NTK). Vzájemná komunikace mezi těmito institucemi je však jen omezená. Proto tvůrci Invenia uspořádali toto setkání administrátorů a vývojářů, jehož hlavním smyslem bylo sdílení zkušeností s provozováním a administrací Invenia, poskytnutí informací o jeho současných funkcích i budoucím vývoji a samozřejmě také navázání nových kontaktů. Program workshopu byl sestaven z přednášek, jejichž témata mohli navrhovat sami účastníci, velký prostor byl vyhrazen pro diskuse.

NTK provozuje v systému Invenio dva digitální repozitáře (repozitář NUŠL a Institucionální digitální repozitář). Hlavním výstupem mé účasti na workshopu mělo být zejména prohloubení znalostí vnitřní struktury i správy Invenia a získání nových praktických poznatků od ostatních administrátorů.

## Časový průběh cesty

K dopravě bylo využito přímé letecké spojení mezi Prahou a Ženevou.

Do Ženevy jsem přijela v neděli 6.5. dopoledne, od pondělí do středy probíhal vždy od 8.30 do 17 až 18 hodin workshop. Odlétala jsem ve čtvrtek 10.5. dopoledne.

## Průběh navštívené akce

Workshop se konal v prostorách Výpočetního centra CERNu. Byl rozdělen na celkem šest tematicky zaměřených bloků, ve kterých jsme si postupně předvedli celou práci se záznamy v Inveniu od jejich vložení do repozitáře přes úpravy až po prezentaci uživatelům, poslední část byla věnována systémové údržbě. Každý blok obsahoval několik přednášek a vedl jej vždy programátor z CERNu zodpovědný za vývoj příslušných funkcí Invenia. Účastníci mohli kdykoliv během přednášek pokládat dotazy vývojářům, ti si také pečlivě zapisovali všechny náměty na možná vylepšení systému.

Následuje shrnutí základních informací z každého bloku.

### Přivítání

Úvodní pondělní blok byl věnován vzájemnému představení všech účastníků workshopu a ukázkám instalací Invenia v jejich domovských institucích. Původ Invenia zde byl hodně znát - téměř polovina z celkem 15 přednášejících pocházela buď přímo z CERNu nebo úzce spřízněných vědecko-technických institucí, zejména různých fyzikálních ústavů (repozitáře CDS, Inspire, IEKP, DESY, JINR, EPFL).

Obecněji zaměřené knihovny byly zastoupeny kromě NTK ještě univerzitními knihovnami v Barceloně a Soluni a také sítí západošvýcarských knihoven RERO. V počátcích je implementace Invenia v africkém institutu IDEP, kde chtějí shromažďovat veškeré africké vysokoškolské kvalifikační práce i další šedou literaturu.

Invenio je jako repozitář dále používáno v evropských projektech OpenAIRE a BlogForever, italská firma nesting.org jej testuje jako úložiště multimediálního obsahu pro své aplikace.

Prakticky všichni účastníci již během této úvodní sekce krátce předvedli své lokální úpravy Invenia (např. změny ve vkládání dokumentů, práci s chráněnými soubory, vlastní skripty pro úpravy záznamů), což bylo dobrým podnětem čilých diskusí během přestávek na kávu.

V rámci této sekce jsem také představila NTK a oba naše digitální repozitáře.

Úvodní blok zakončil vedoucí vývoje Tibor Šimko **stručným přehledem současné struktury a funkcí Invenia** s výhledem do blízké budoucnosti. Díky novému způsobu vývoje se můžeme těšit na novou minor verzi každé tři až čtyři měsíce - nejbližší verze 1.1 je již prakticky hotová a její vydání je naplánováno na červenec. V delším časovém horizontu proběhnou hluboké změny technologií, na kterých je postavena většina modulů: aby bylo možné používat různé typy databází, bude přidán ORM (konkrétně SQLAlchemy), aplikace bude vyvíjena ve frameworku Flask a oddělení prezentační vrstvy zajistí šablonovací systém Jinja2.

## Získávání a vkládání záznamů do Invenia

Nejpodrobněji byly rozebrány moduly WebSubmit a BibConvert, kde se vytvářejí a zpracovávají formuláře pro přímé vkládání záznamů do repozitáře. Postupně jsme prošli celý **proces, který probíhá při vložení záznamu**, a který je v současné době velmi komplikovaný. Jeho složitost přiměla hned několik administrátorů k lokálním úpravám těchto modulů, také v NTK používáme pro některé části zpracování a konverze záznamů vlastní skript.

Následovala přednáška o **harvestování** OAI zdrojů, kde byly předvedeny zajímavé možnosti dodatečného zpracování a filtrování získaných záznamů - v repozitáři Inspire takto například z dokumentů získávají a zvlášť ukládají citace a grafy.

Neméně zajímavý byl příspěvek o **harvestování zdrojů, které neodpovídají protokolu OAI-PMH**. Stačí, když má takový zdroj jakoukoliv pevně danou strukturu, která se dá namapovat do formátu MARCXML. Takové zdroje lze teoreticky zpracovávat již dnes, nová verze 1.1 ale obsahuje nástroje, které podstatně usnadní automatizaci tohoto procesu. Jako příklad nám vývojař Samuele Kaplun předvedl harvestování francouzských televiziních programů.

Od verze 1.0 Invenio umožňuje také **plně automatizované nahrávání** metadat (ve formátu MARCXML) i dokumentů do repozitáře pomocí http metody POST a také **dávkové nahrávání** metadat i souborů přes administrátorské webové rozhraní (přitom ale musí být splněno několik podmínek, například názvy souborů musí přesně odpovídat číslům záznamů, ke kterým patří).

Poslední pondělní příspěvky se týkaly specifik **práce s multimediálními soubory** (například právě na fotogalerie se hodí výše zmíněné automatické či hromadné nahrávání, chystá se modul pro převádění videí do různých formátů a rozměrů - původní velikost se obvykle nehodí pro web) a představení nového modulu **BibIngest**, který by měl zobecnit ukládání metadat a umožnit například použití různých formátů (json) a různých databází (i nerelačních), zatím je však ve velmi rané fázi vývoje.

## Zpracování a úpravy záznamů v repozitáři

Viděli jsme podrobný návod na práci se záznamy ve webovém administrátorském rozhraní: kromě obvyklé editace jednoho záznamu jsou možné také **hromadné úpravy záznamů** na základě libovolného vyhledávacího kritéria (například přidání pole do všech záznamů, které se nachází v určitých sbírkách) nebo otevření dvou záznamů vedle sebe a přenos hodnot polí z jednoho záznamu do druhého (hodí se to například při práci s potenciálními duplicitami).

Z editorského rozhraní je dostupná také funkce **„holding pen"** - nově vkládané záznamy mohou být ukládány do speciální fronty, kde čekají na kontrolu administrátora, a až po ní se přesunují do repozitáře.

Hodně času jsme věnovali **práci s dokumenty**: k záznamu je můžeme připojit buď jen odkazem v poli 8564_u, nebo s nimi přímo manipulovat pomocí tzv. FFT tagů, kde jsou uloženy informace o cestě k souboru, jeho popis, omezení přístupu, verze a mnoho dalších. Invenio dále obsahuje mocné nástroje k získávání statistik o souborech, pro konverzi mezi různými formáty (včetně např. OCR a vytvoření textové vrstvy nad PDF) či k extrakci metadat (např. EXIF z fotografií).

Posledním větším úterním tématem byla správa autorit. V současnosti lze k tomuto účelu využít knowledge bases (součást Invenia, jde o jednoduché seznamy nebo key-value úložiště) nebo speciální záznamy (v repozitáři Inspire jsou sbírky "konference" a "instituce" s kompletními daty, z běžných záznamů sem pak vede odkaz). Chystá se modul BibAuthority, možná bude k dispozici už ve verzi 1.2 (dle plánu cca za půl roku).

## Prezentace záznamů uživatelům

Poslední den workshopu jsme začali vyhledáváním: předvedli jsme si pokročilou vyhledávací syntaxi (Invenio umí vyhledávat i regulární výrazy) a nízkoúrovňová vyhledávací API pro skriptování.

V příštích verzích bude ve vyhledávání mnoho změn a vylepšení, například sjednocení významu jednoduchých a dvojitých uvozovek (současný stav je pro uživatele docela matoucí), nové možnosti řazení záznamů podle jejich kvality - oblíbenosti u uživatelů (podobně jako u Googlu se bude brát v potaz počet zobrazení detailu záznamu, počet stažení jeho souborů a podobně), celkové zrychlení vyhledávání v metadatech i plných textech díky novým technologiím, webové rozhraní pokročilého vyhledávání bude zjednodušeno a zároveň do něj přibudou facety.

Do Invenia je možné připojit i externí sbírky, a to dvěma způsoby: volně, kdy lze v takových sbírkách snadno vyhledávat souběžně s vyhledáváním v domovském repozitáři, anebo jako tzv. „hosted" sbírky, které se pak navenek tváří, jako by byly součástí repozitáře a lze je také shodně používat (například ukládat si záznamy z nich do košíku, nastavovat emailová upozornění na nové položky a podobně).

Velmi zajímavé jsou možnosti propojení Invenia s webovými prohlížeči a jejich pluginy pro správu citací a vyhledávání, např. Zotero, LibX, OpenSearch.

Dále jsme se v této části zabývali formátováním záznamů a řízením přístupových práv, stručně byl předveden výpůjční modul.

## Údržba systému a programování

Poslední blok byl zaměřen ryze technicky, od jednotlivých modulů Invenia jsme se přesunuli na nižší úroveň a věnovali se systémům, na kterých je celé Invenio postaveno.

Velmi užitečná byla přednáška o tom, kde hledat chybu, když se něco pokazí. Mnoho procesů, které v Inveniu běží, zaznamenává svou činnost do různých logů, téměř všechny funkce se také dají spustit s parametrem, který zajistí podrobné výpisy o jejich průběhu.

Dále jsme se věnovali architektuře systému a možnostem nastavení a sledování výkonu jednotlivých komponent (OS, Apache, MySQL, WSGI).

Vývoj Invenia má na starosti především tým v CERNu, který zároveň vítá zapojení ostatních členů komunity. Ta je v tomto směru docela aktivní, například jen za minulý rok (2011) přispělo se svým kódem téměř 40 lidí. Při takovém způsobu vývoje je velmi důležité, aby byl zdrojový kód dobře zorganizovaný a aby byly nastaveny mechanismy pro kontrolu kvality. Kód Invenia je spravován ve verzovacím systému Git, v několika hlavních větvích zahrnujících starší verze i zatím nehotové a experimentální části. Velký důraz je kladen na testování, příspěvky od komunity by také měly splňovat několik požadavků, které byly uvedeny v závěru této přednášky.

Workshop uzavřel podrobnější pohled na již zmíněné nové technologie, které se chystají do verze 2.0, zejména na SQLAlchemy. Tento ORM je již nyní částečně implementován a otestován na jednom z modulů.

## Shrnutí

Workshop mi poskytnul velmi dobrý přehled o současných i plánovaných funkcích Invenia, dozvěděla jsem se mnoho tipů ke správě systému, získala jsem nové podněty k vylepšení instalací Invenia v NTK a navázala kontakty s ostatními administrátory.

Přínos cesty obzvlášť vyniká ve světle skutečnosti, že mnoho funkcí Invena je zdokumentovaných jen zčásti nebo dokonce vůbec, takže bych se informace nabyté během tří dnů workshopu jinak dozvídala podstatně déle a nahodileji.

## Dokumentace

Podrobný program workshopu je přílohu této cestovní zprávy, všechny prezentace jsou k dispozici na stránce http://indico.cern.ch/conferenceDisplay.py??confId=183318

## Anotace ZC

Náplní cesty byla účast na mezinárodním semináři Invenio User Group Workshop, který se konal 7.-9.5. 2012 ve Výpočetním centru CERNu v Ženevě. Workshop byl rozdělen do šesti tematických bloků, které vedli jednotliví vývojáři Invenia. Postupně předvedli celou práci se záznamy od jejich vložení do repozitáře, přes úpravy až po prezentaci uživatelům, poslední blok byl věnován systémové údržbě a způsobu vývoje nových komponent. Důležitou součástí workshopu byly také vzájemné diskuse všech účastníků. Výstupem z workshopu je mnoho nových praktických poznatků, které budou přímo využity při správě obou digitálních repozitářů NTK.

# Invenio User Group Workshop 2012

# Monday 07 May 2012

## Welcome : roundtable session of all present Invenio instances - 513-1-024 (08:45-12:00)

**- Conveners: Mr. Le Meur, Jean-Yves**

| time | title | presenter |
|---|---|---|
| 08:45 | Welcome to CERN | FOSTER, David - Deputy Head of IT Department<br>SMITH, Tim |
| 09:00 | Invenio @ CERN | LE MEUR, Jean-Yves |
| 09:10 | Invenio @ CDS | GENTIL-BECCOT, Anne |
| 09:20 | Invenio @ INSPIRE | LAVIK, Jan Age |
| 09:30 | Invenio @ Uni-Karlsruhe | RATNIKOVA, Natalia |
| 09:40 | Invenio @ AUTH | THEODOROPOULOS, Theodoros |
| 09:50 | Invenio @ M9 | BACCHI, Cristian |
| 10:00 | Invenio @ nesting.org | GROSSO, Mattia |
| 10:10 | Coffee break | |
| 10:30 | Invenio @ EPFL | FAVRE, Gregory |
| 10:40 | Invenio @ OpenAIRE | NIELSEN, Lars Holm |
| 10:50 | Invenio @ UAB | JORBA, Ferran |
| 11:00 | Invenio @ DESY, GSI, FZ Juelich and RWTH Aachen | HESSELBACH, Stefan |
| 11:10 | Invenio @ Techlib | SLOUKOVA, Jana |
| 11:20 | Invenio @ JINR | MUSULMANBEKOV, Genis |
| 11:30 | Invenio @ RERO | MARIéTHOZ, Johnny |
| 11:40 | Invenio @ Blogforever | KASIOUMIS, Nikos |
| 11:50 | Invenio @ UNIDEP (Dakar) | GUEDEGBE, Eric |

## Status of INVENIO: overview of versions and features - 513-1-024 (13:00-13:30)

***INVENIO Latest Master - New features***

**- Conveners: Dr. Simko, Tibor**

| time | title | presenter |
|---|---|---|
| 13:00 | Invenio Status: Versions and Features | SIMKO, Tibor |

## INGESTION Modules: everything you want to learn about ingestion techniques in Invenio - 513-1-024 (13:30-18:00)

***Ingestion techniques in Invenio, covering WebSubmit, BibHarvest, BatchUploader, etc. Integration with other systems such as Indico.***

**- Conveners: Mr. Caffaro, Jerome**

| time | title | presenter |
|---|---|---|
| 13:30 | Ingestion - Introduction | CAFFARO, Jerome |
| 13:35 | Configuration of web-forms for submission (WebSubmit) | CAFFARO, Jerome |
| 14:30 | OAI-Harvesting (BibHarvest) | LAVIK, Jan Age |
| 15:20 | Non OAI Harvesting / Ingestion | KAPLUN, Samuele |
| 15:40 | Coffee break | |
| 16:00 | Integration with Indico | GONZALEZ LOPEZ, Jose Benito |
| 16:30 | Batch ingestion | MARTIN MONTULL, Javier<br>KAPLUN, Samuele |
| 16:50 | Ingestion Store, OAIS | KASIOUMIS, Nikos |
| 17:05 | Multimedia content handling | CAFFARO, Jerome |

## Hands-on session - 513-1-023 (15:00-17:30)

***Invenio expert available during the workshop for discussions and hands-on activities***

## Dinner in Geneva at http://g.co/maps/tj8jk - (19:00-22:35)

***Social event: dinner in Geneva, at "Chez ma Cousine", close to the train station***

***See http://g.co/maps/679ue***

# Tuesday 08 May 2012

## PROCESSING: indexing, ranking, sorting, scheduling and more internal Invenio processes - 513-1-024 (08:30-10:00)

**- Conveners: Dr. Simko, Tibor**

| time | title | presenter |
|---|---|---|
| 08:30 | Invenio Daemons: Scheduler, Classification, Indexing, Ranking | SIMKO, Tibor |
| 09:15 | New sorting facility: BibSort | MARIAN, Ludmila |
| 09:30 | New combined distributed ranking facility | MARIAN, Ludmila |
| 09:45 | Using external ranking tools: Solr, Xapian | GLAUNER, Patrick Oliver |

## Hands-on session - 513-1-023 (10:00-18:00)

***Invenio expert available during the workshop for discussions and hands-on activities***

## CURATION: Invenio facilities to edit, correct, enrich ingested content - 513-1-024 (13:30-17:35)

***Review of the modules relative to the Curation processes in Invenio***

**- Conveners: Mr. Kaplun, Samuele**

| time | title | presenter |
|---|---|---|
| 13:30 | Records Editing: BibEdit, MultiEdit | MARTIN MONTULL, Javier<br>DEIANA, Alessio |
| 14:30 | File Management with BibDocFile | KAPLUN, Samuele |
| 15:30 | Knowledge bases | SIMKO, Tibor |
| 15:45 | Coffee break | |
| 16:05 | Using authority records | SIMKO, Tibor |
| 16:20 | Author management: BibAuthorId | CARLI, Samuele |
| 16:55 | Records deduplication: BibMatch | SIMKO, Tibor |
| 17:10 | BibClassify | SIMKO, Tibor |

# Wednesday 09 May 2012

## DISSEMINATION: how to best make Invenio content available - 513-1-024 (08:30-12:30)

***Review of all modules relative to the dissemination of the content to users and other systems.***

**- Conveners: Marian, Ludmila**

| time | title | presenter |
|---|---|---|
| 08:30 | Searching: syntax, APIs | MARIAN, Ludmila<br>COSTA, Flavio |
| 09:00 | Collection management | MARIAN, Ludmila<br>KASIOUMIS, Nikos |
| 09:25 | Collaborative tools: baskets, alerts, groups, comments | KASIOUMIS, Nikos<br>CAFFARO, Jerome |
| 09:55 | Coffee break | |
| 10:10 | Authentication and access control: WebAccess | KAPLUN, Samuele |
| 10:55 | Circulation and Holdings: BibCirculation | GARCIA LLOPIS, Jaime |
| 11:25 | Record Export | CAFFARO, Jerome |
| 12:10 | Collecting statistics: WebStat | SIMKO, Tibor |

## Hands-on session - 513-1-023 (10:00-17:30)

***Invenio expert available during the workshop for discussions and hands-on activities***

## Invenio Best Practices: System configurations, contibuting to the codebase, etc. - 513-1-024 (13:30-16:30)

***Developing code, contributing code, system performance and configuration, operational troubleshooting.***

**- Conveners: Dr. Simko, Tibor**

| time | title | presenter |
|---|---|---|
| 13:30 | Best System Practices: architecture, performance, profiling, tuning | SIMKO, Tibor |
| 14:00 | Best Development Practices: branches, tickets, code kwalitee, local developments | SIMKO, Tibor |
| 15:00 | Operational alarms, task queue, log inspection, troubleshooting | KAPLUN, Samuele |
| 15:20 | Coffee break | |
| 15:40 | Customizing templates: WebStyle | CAFFARO, Jerome |
| 16:00 | Database Independence: SQLAlchemy | KUNCAR, Jiri |

## Wrap-up Workshop - 513-1-024 (16:30-17:30)

**- Conveners: Mr. Le Meur, Jean-Yves**