Sustainable linguistic repository built on DSpace

Jozef Mišutka
LINDAT/CLARIN dev team

Institute of Formal and Applied Linguistics
Charles University in Prague, Czech Republic

LINDAT/CLARIN

- Common Language Resources and Technology Infrastructure
- LINDAT/CLARIN = CLARIN centre in the Czech Republic
- one of the offered services is a digital repository for sharing linguistic data and tools
- must meet requirements for interoperability, AAI, preservation and sustainability

Let us improve DSpace together

- we want to take an **active** part in merging our changes if there is an **active** developer interest
- links to more information do not need to reflect the absolute latest development

Our contributions and know-how

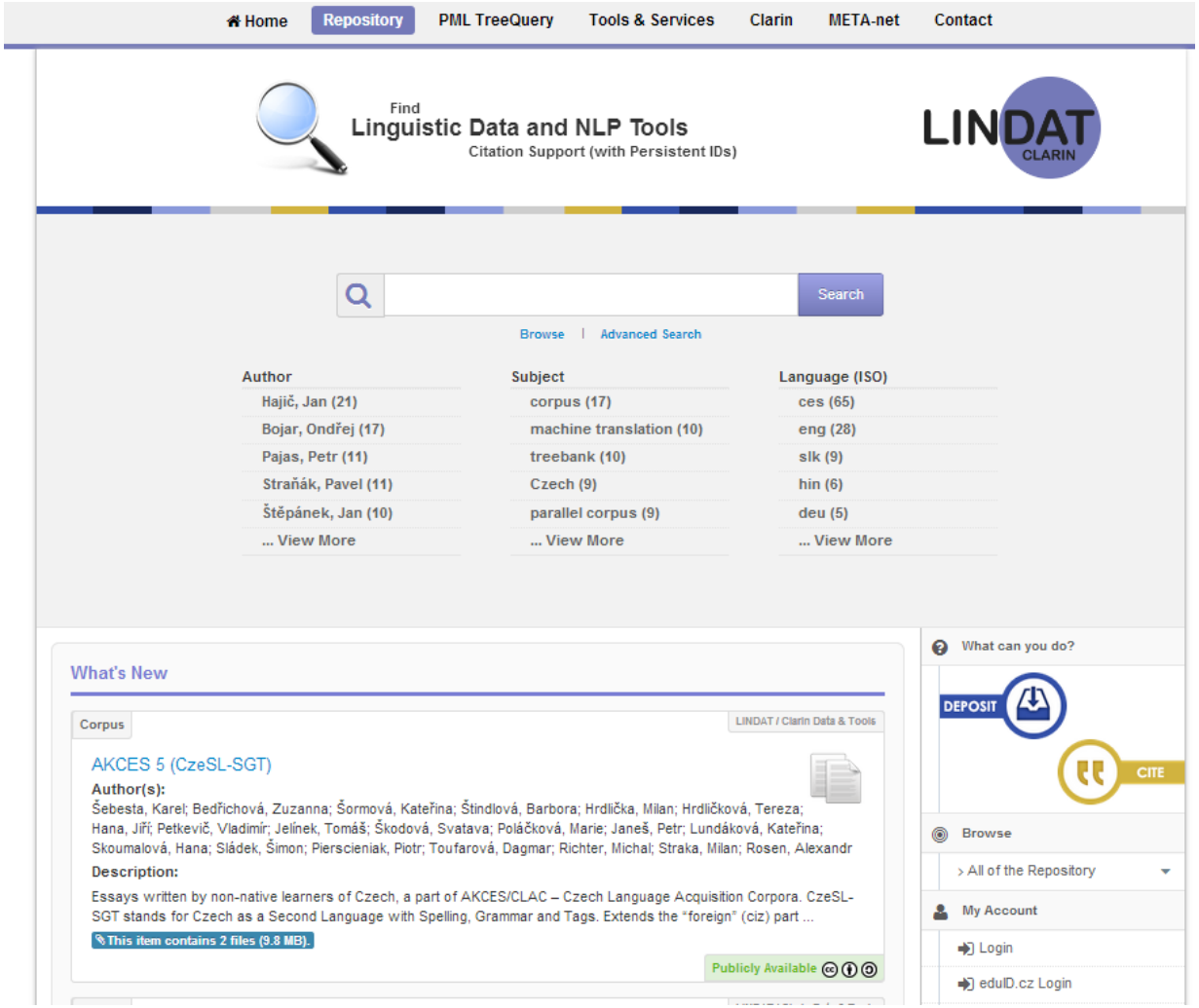
- user interface using bootstrap
- control panel, weekly reports, logger
- submission workflow, ACL fields, drag’n’drop
- data protection law
- global user base - AAI
- licensing model
- PID - handle subsystem
- backup and replication
- metadata, export formats and QA

User Interface

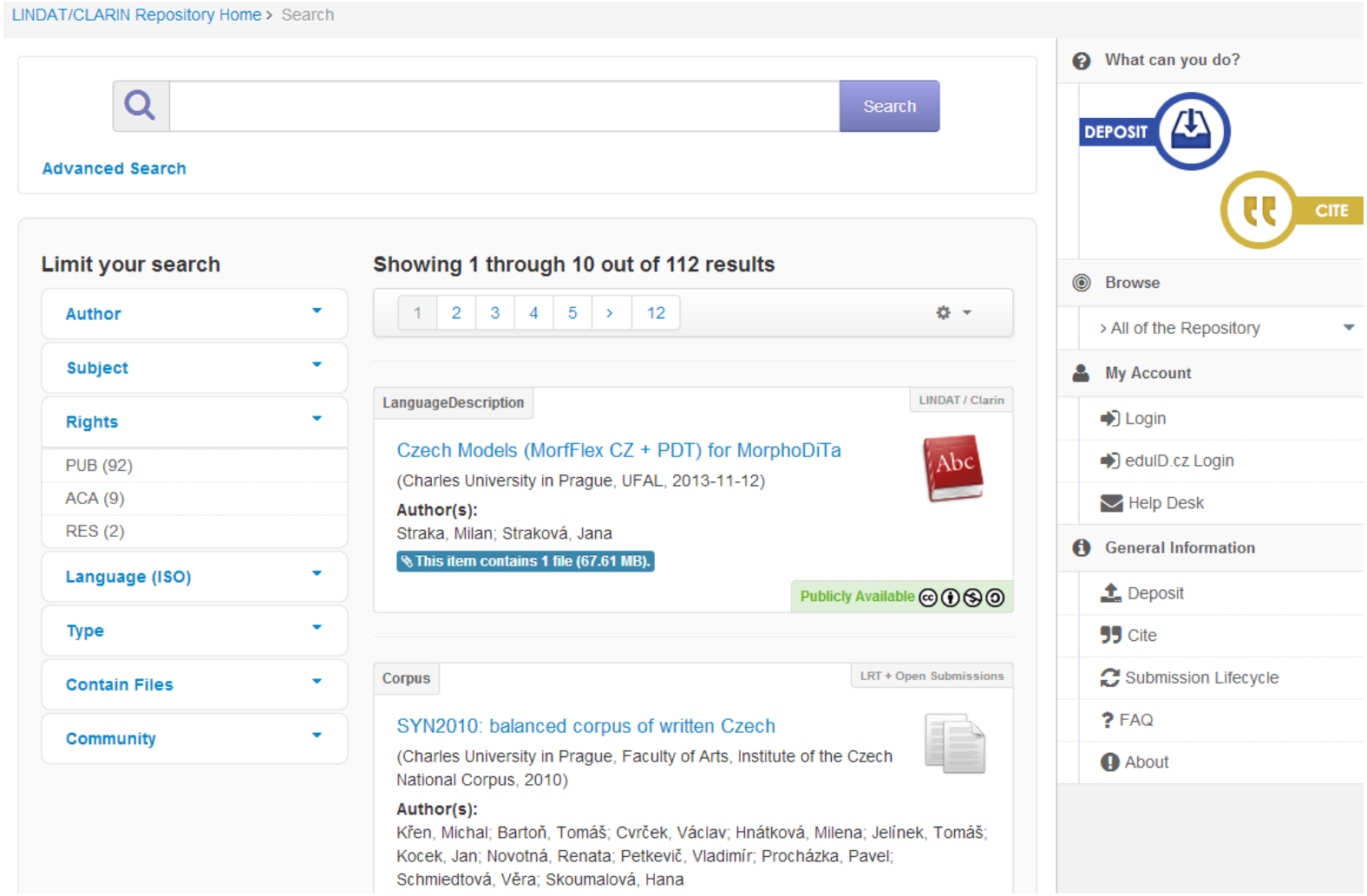
- new XMLUI theme built on **bootstrap** library
  - UFALHome - home page theme
  - UFAL - general theme
- using project specific header/footer
- pages can include additional libraries using page metadata “include-library”
- deployed in January 2014

https://svn.ms.mff.cuni.cz/redmine/projects/dspace-modifications/repository/revisions/master/show/sources/dspace/modules/xmlui/src/main/webapp/themes

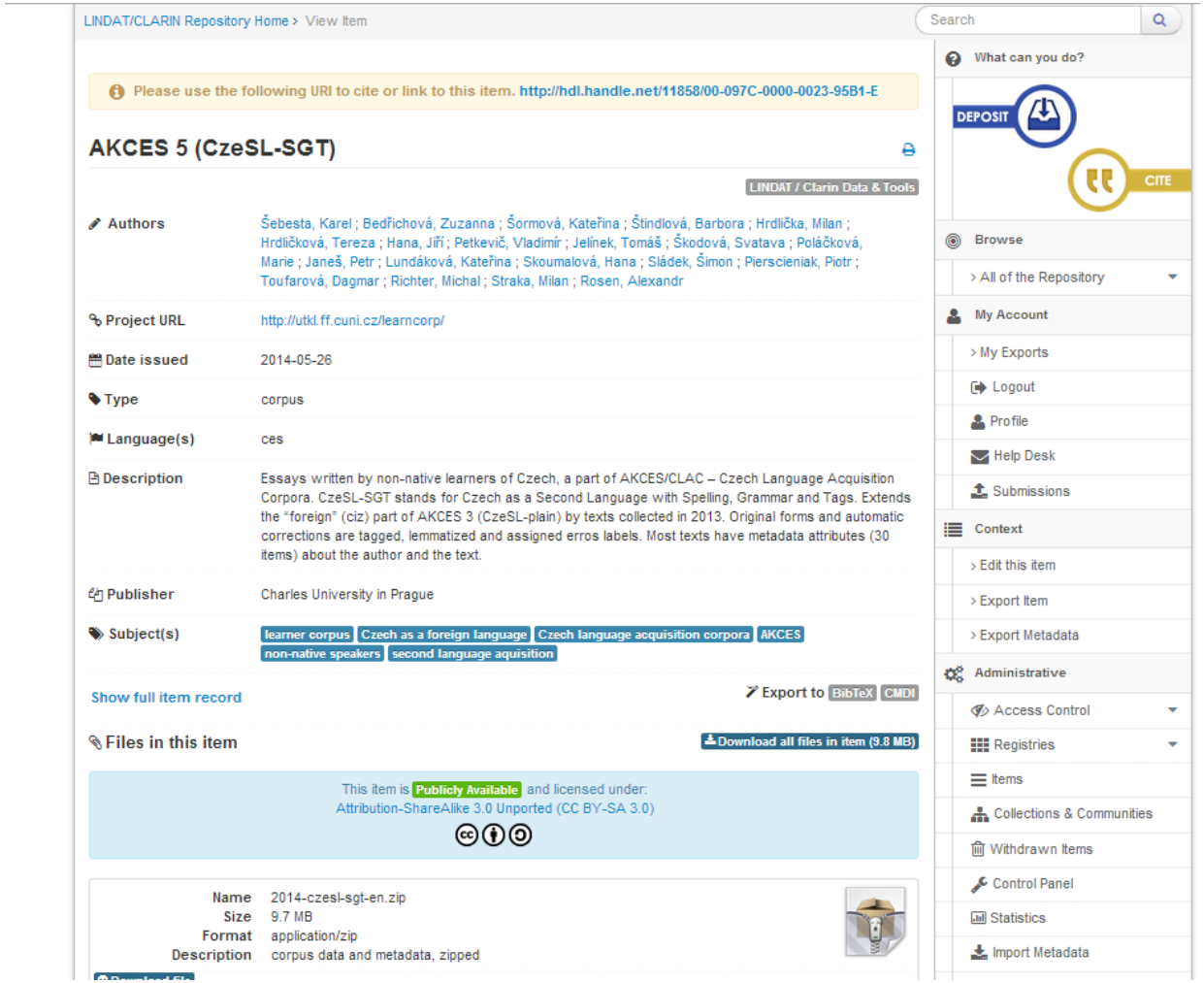
User Interface - home page theme



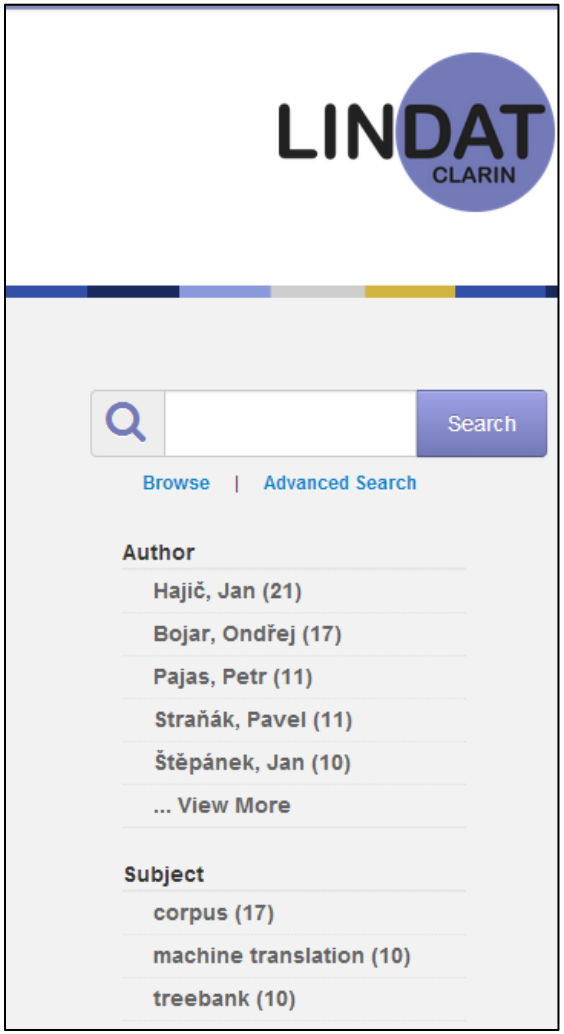
User Interface - searching

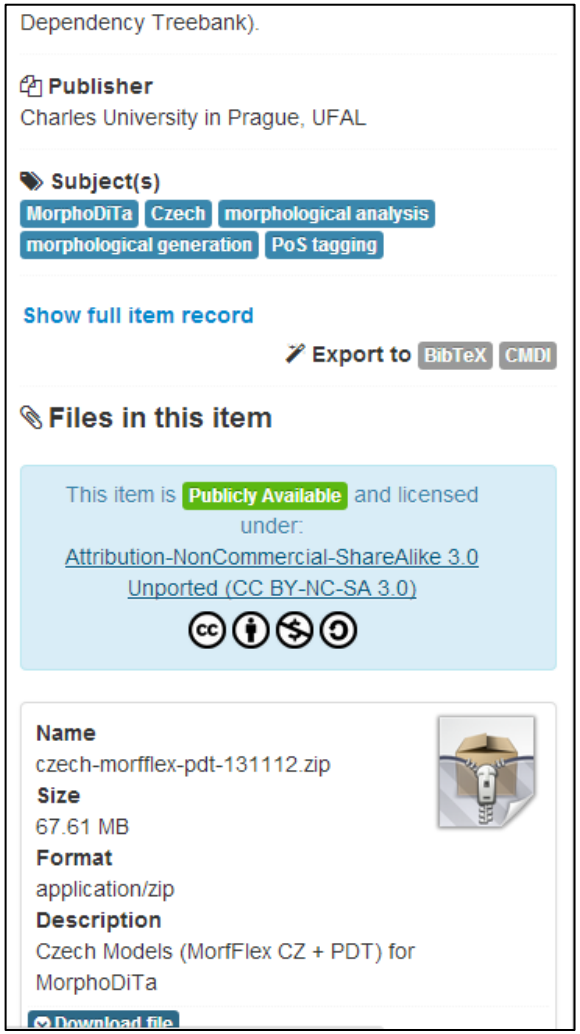


User Interface - item view



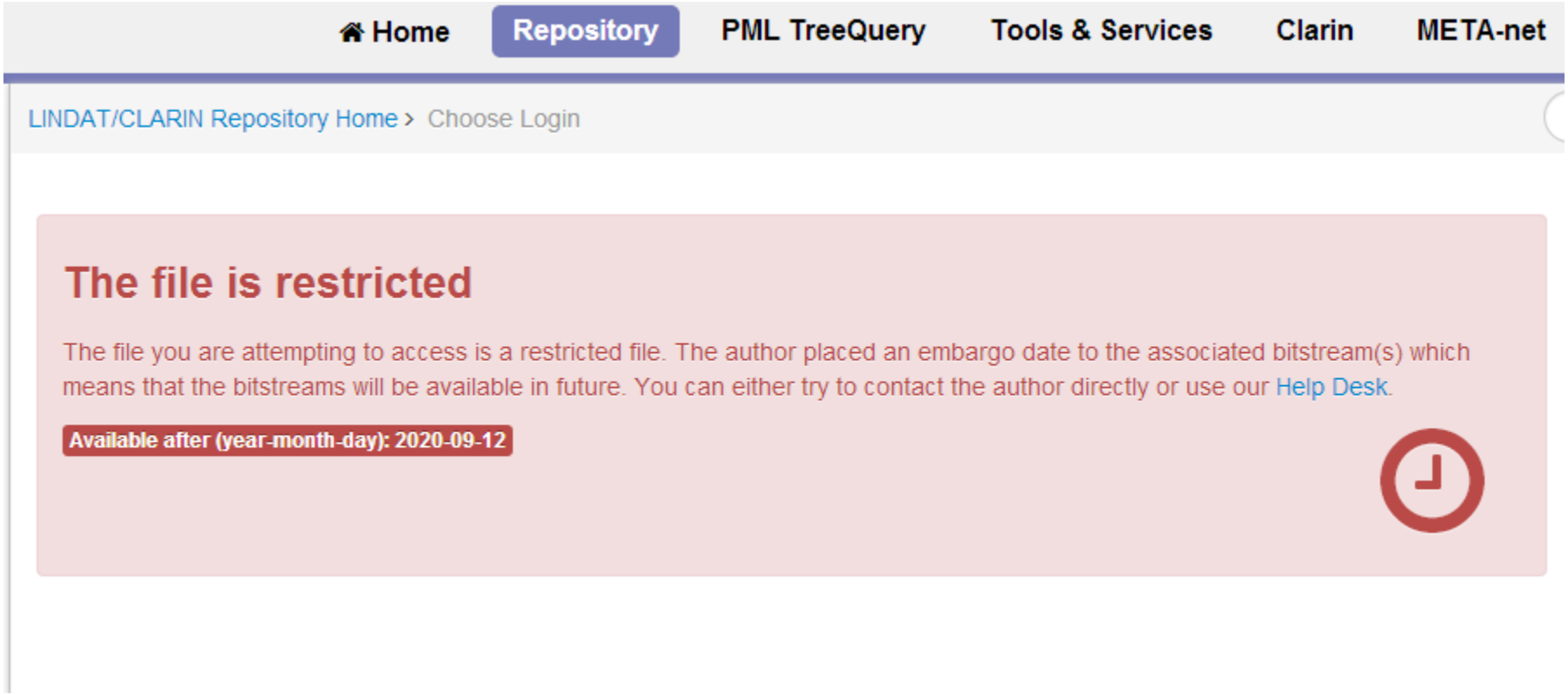
User Interface - mobile



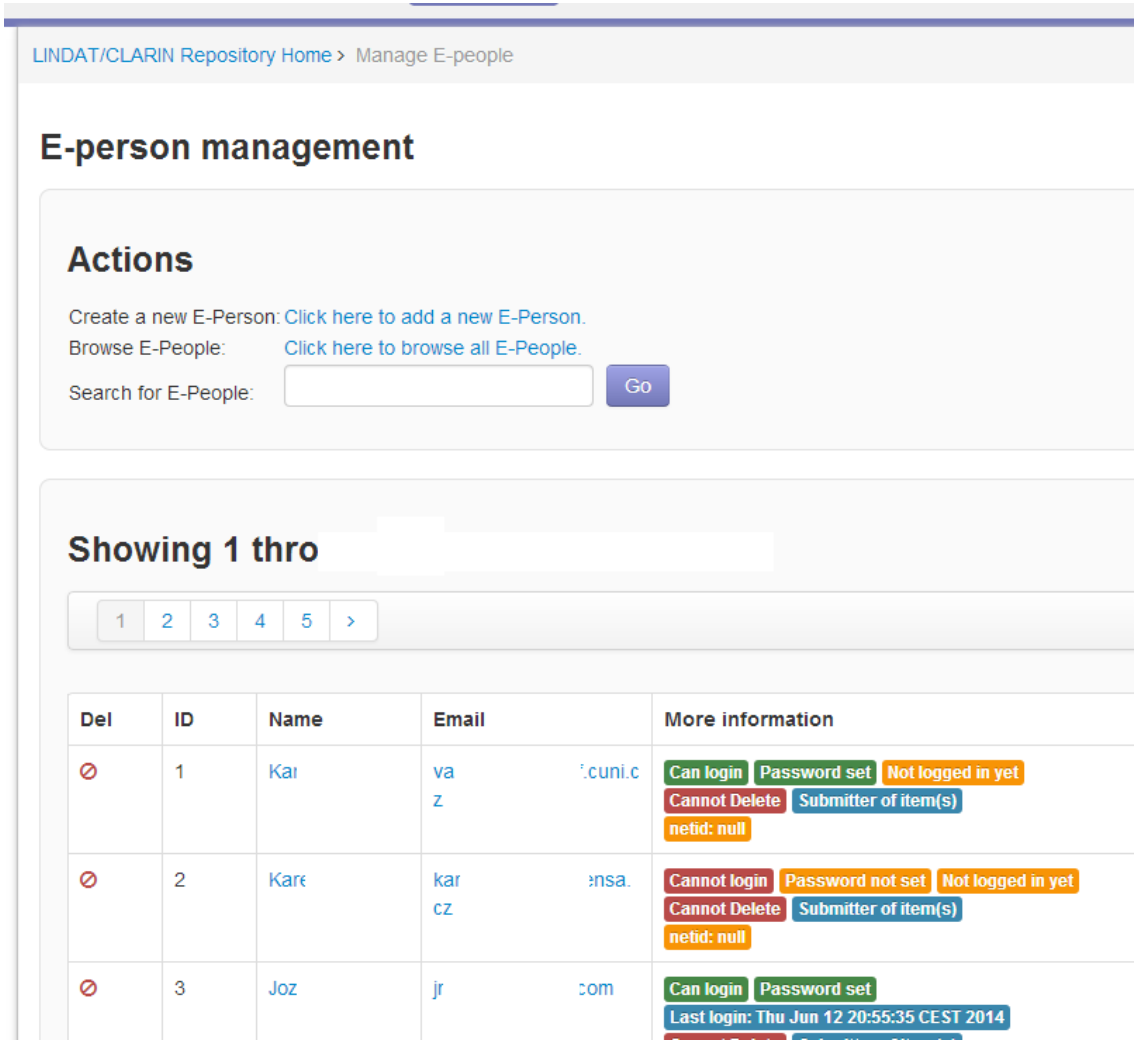


User Interface - replaced by

User Interface - embargo



User Interface - eperson list



Control Panel

- truly **control** and inspect DSpace
- no more if-else and hard to debug code
- enumerate classes representing tabs in controlpanel.cfg
- extensible and easily configurable
- support for asynchronous “checks”

https://svn.ms.mff.cuni.cz/redmine/projects/dspace-modifications/wiki/Controlpanel

Control Panel implementation

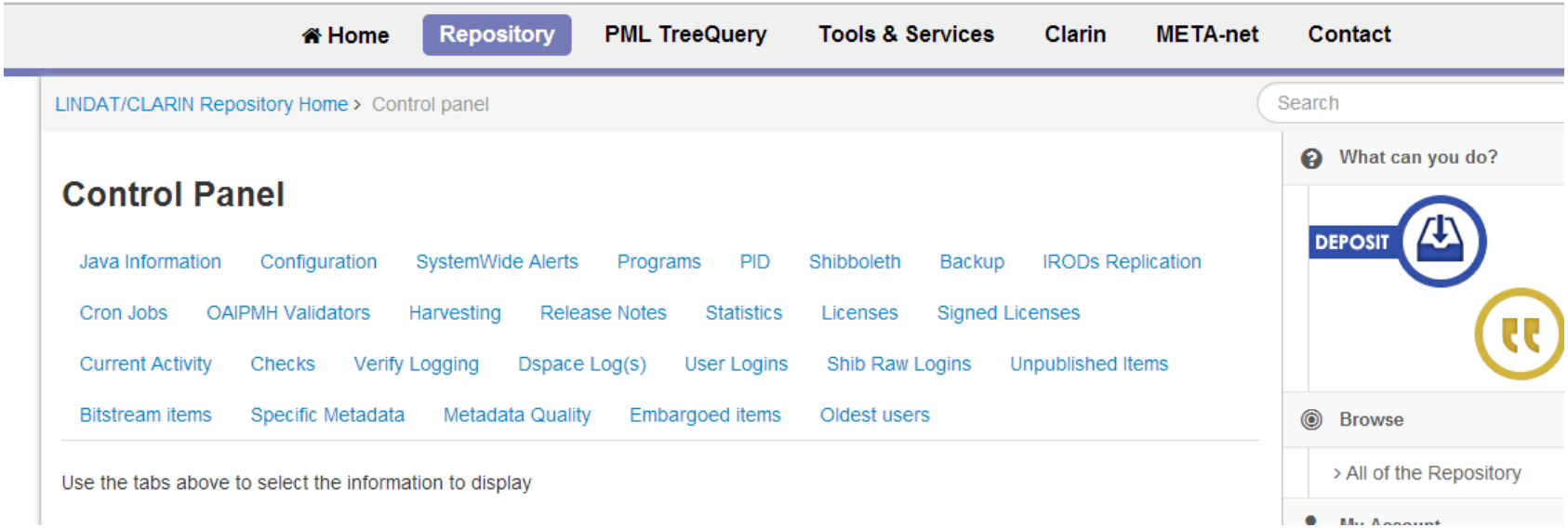
- controlpanel.cfg

```
controlpanel.tabs = Test Panel
plugin.named…..administrative.ControlPanelTab = \
        cz.cuni.mff.ufal.ControlPanelTestTab = Test Panel
```

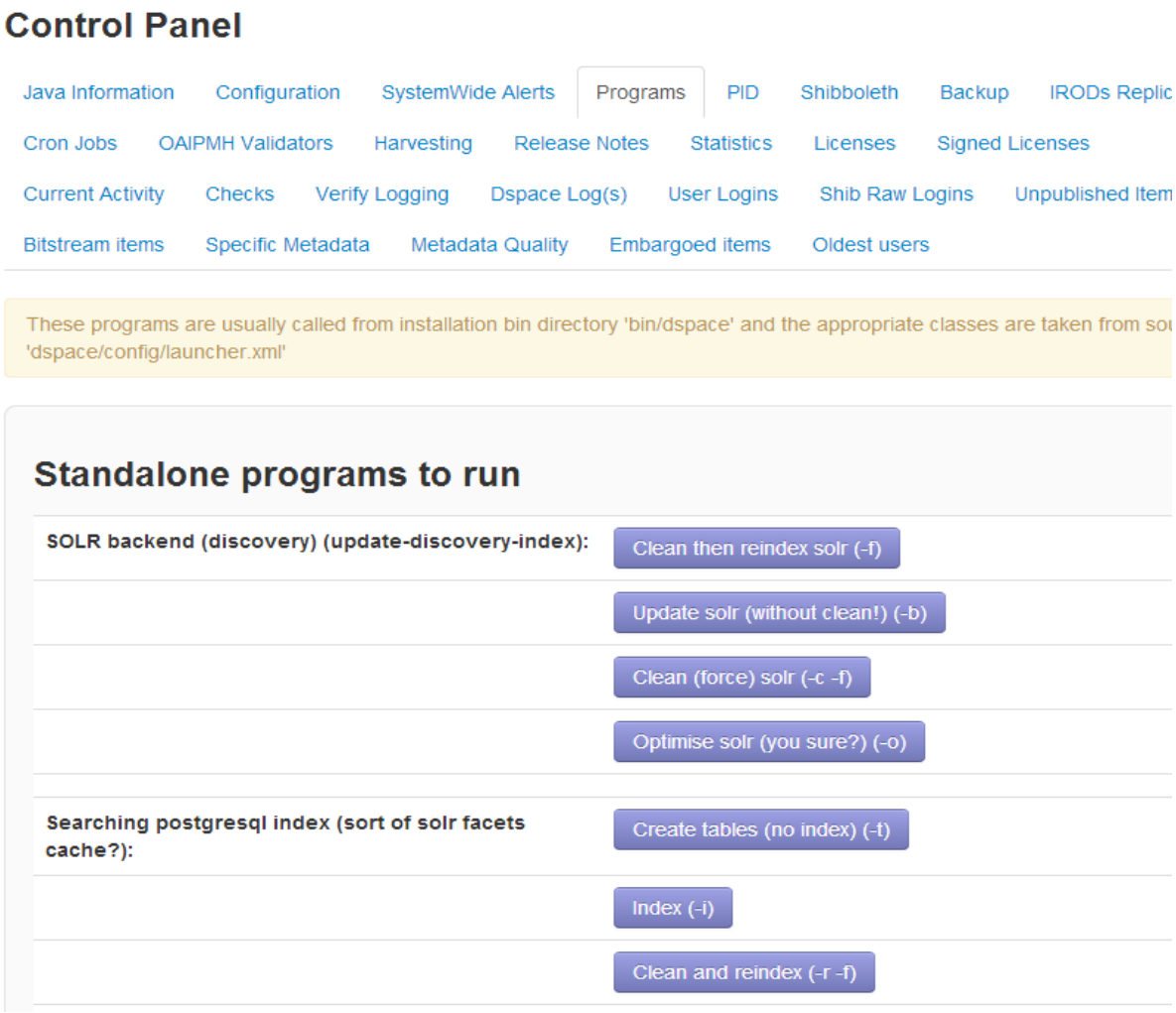
- ControlPanelTestTab.java

```
public class ControlPanelTestTab extends AbstractControlPanelTab {
    @Override
    public void addBody(Map objectModel, Division div) throws WingException {
        div.addPara("This is a tab.");
    }
```

Control Panel



Control Panel - programs



Control Panel - backup

Control Panel

Java Information | Configuration | SystemWide Alerts | Programs | PID | Shibboleth | Backup | IRODs Replication

Cron Jobs | OAIPMH Validators | Harvesting | Release Notes | Statistics | Licenses | Signed Licenses

Current Activity | Checks | Verify Logging | Dspace Log(s) | User Logins | Shib Raw Logins | Unpublished Items

Bitstream items | Specific Metadata | Metadata Quality | Embargoed items | Oldest users

Available Backups

Returned stdout:

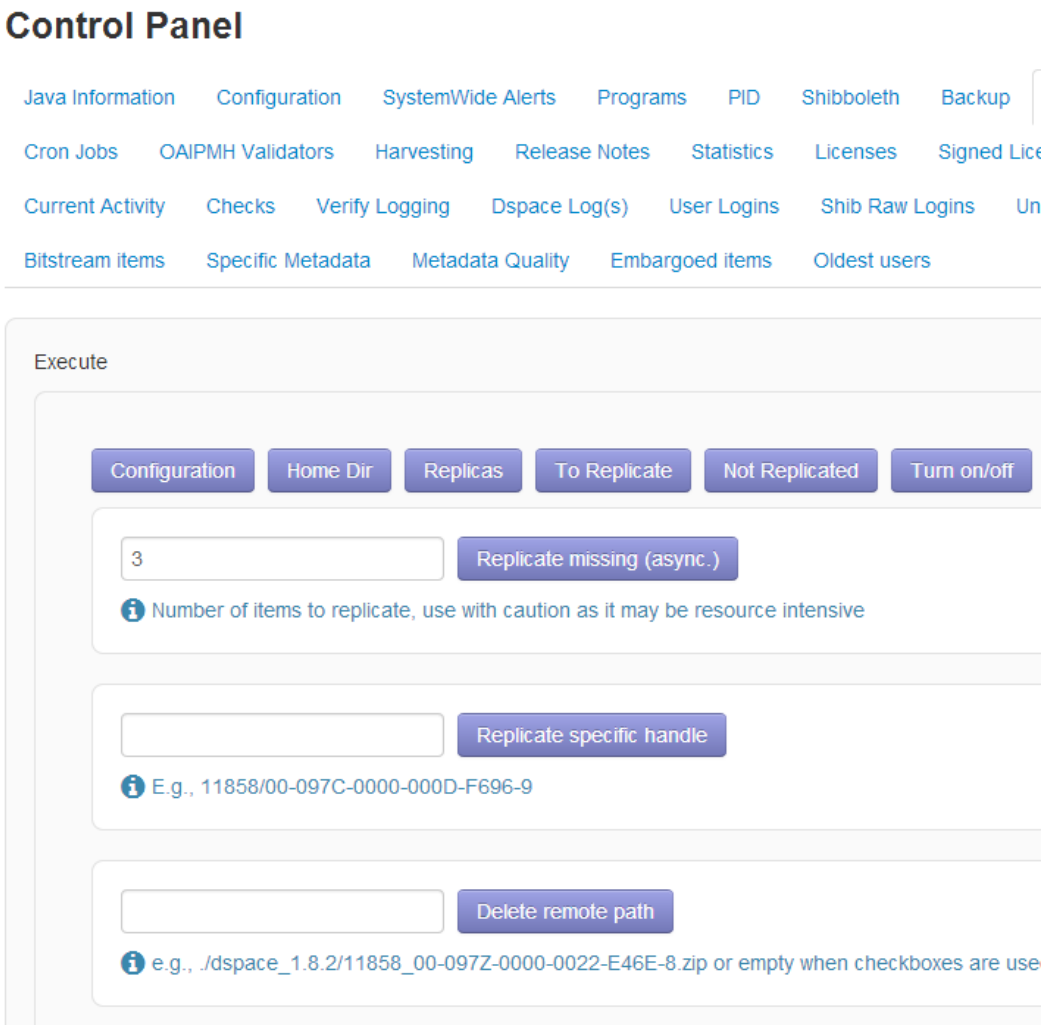
backup2l v1.5 by Gundolf Kiefer

Summary

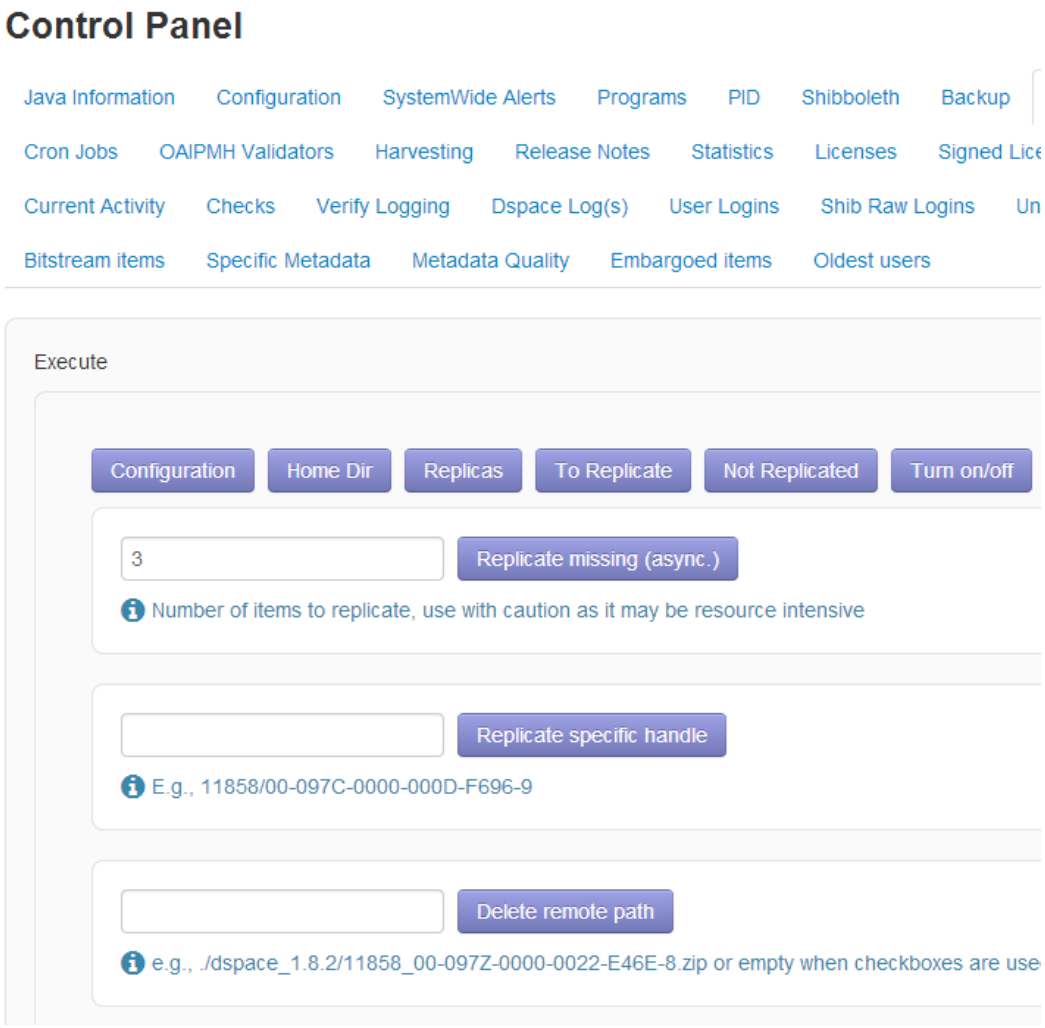
=======

	Backup	Date	Time	Size	Skipped	Files+D	New	Obs.	Err.
	all.1	2014-06-03	02:00	111.7G	0	23078	23078	0	72
	all.11	2014-06-04	02:00	21.1M	0	23083	121	116	0
	all.12	2014-06-05	02:00	21.2M	0	23090	117	110	0
	all.13	2014-06-06	02:00	21.2M	0	23083	96	103	0
	all.14	2014-06-07	02:00	21.2M	0	23114	114	83	0

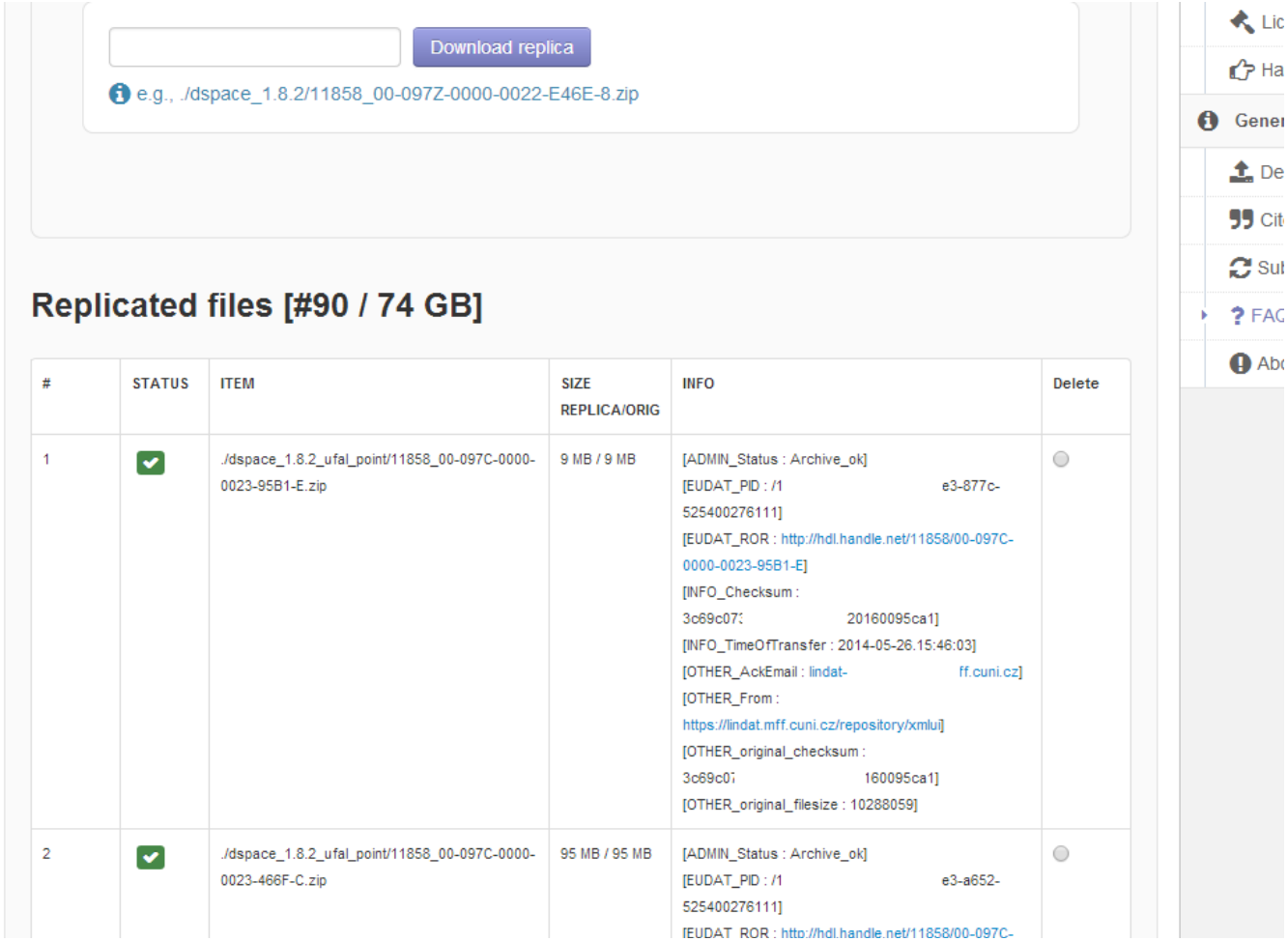
Control Panel - replication



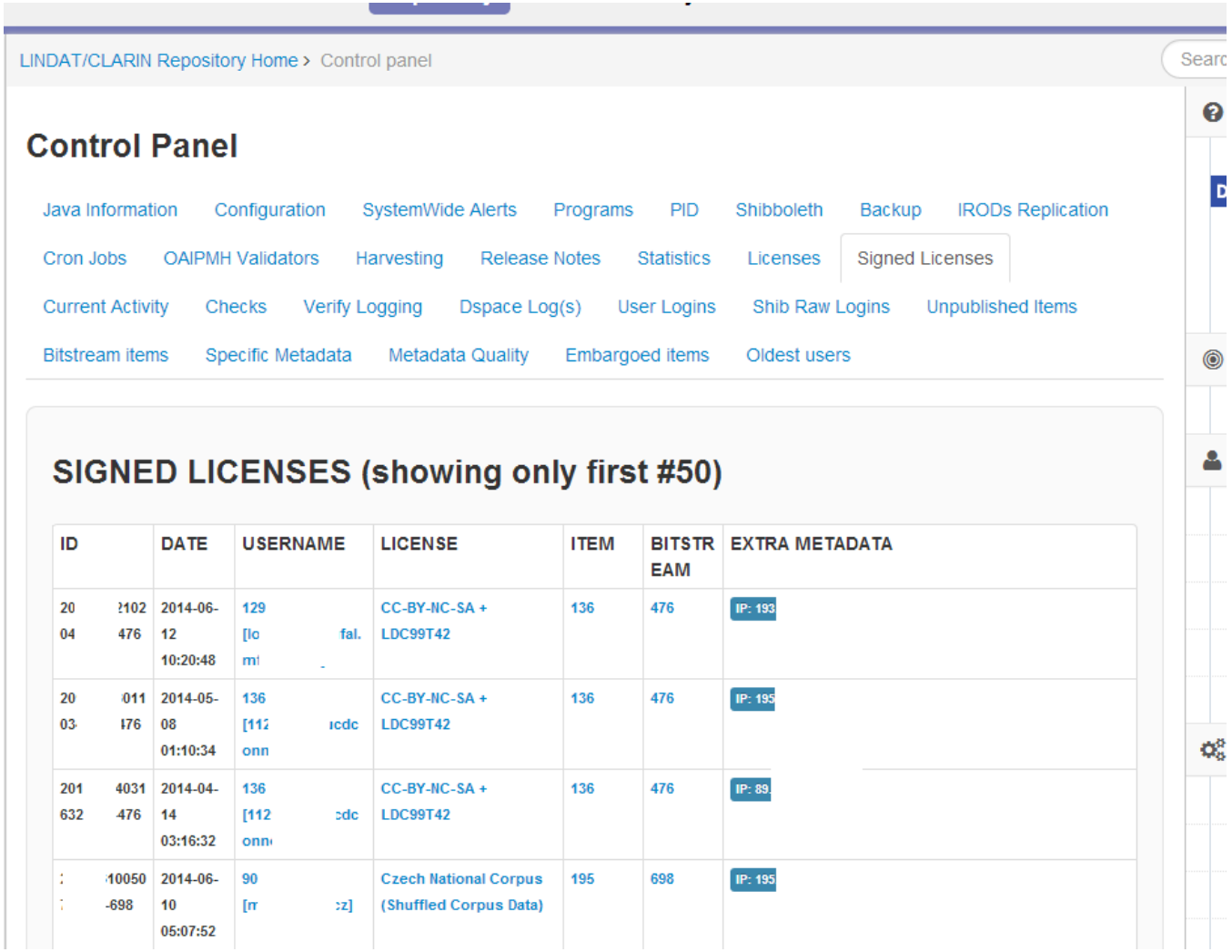
Control Panel - replication



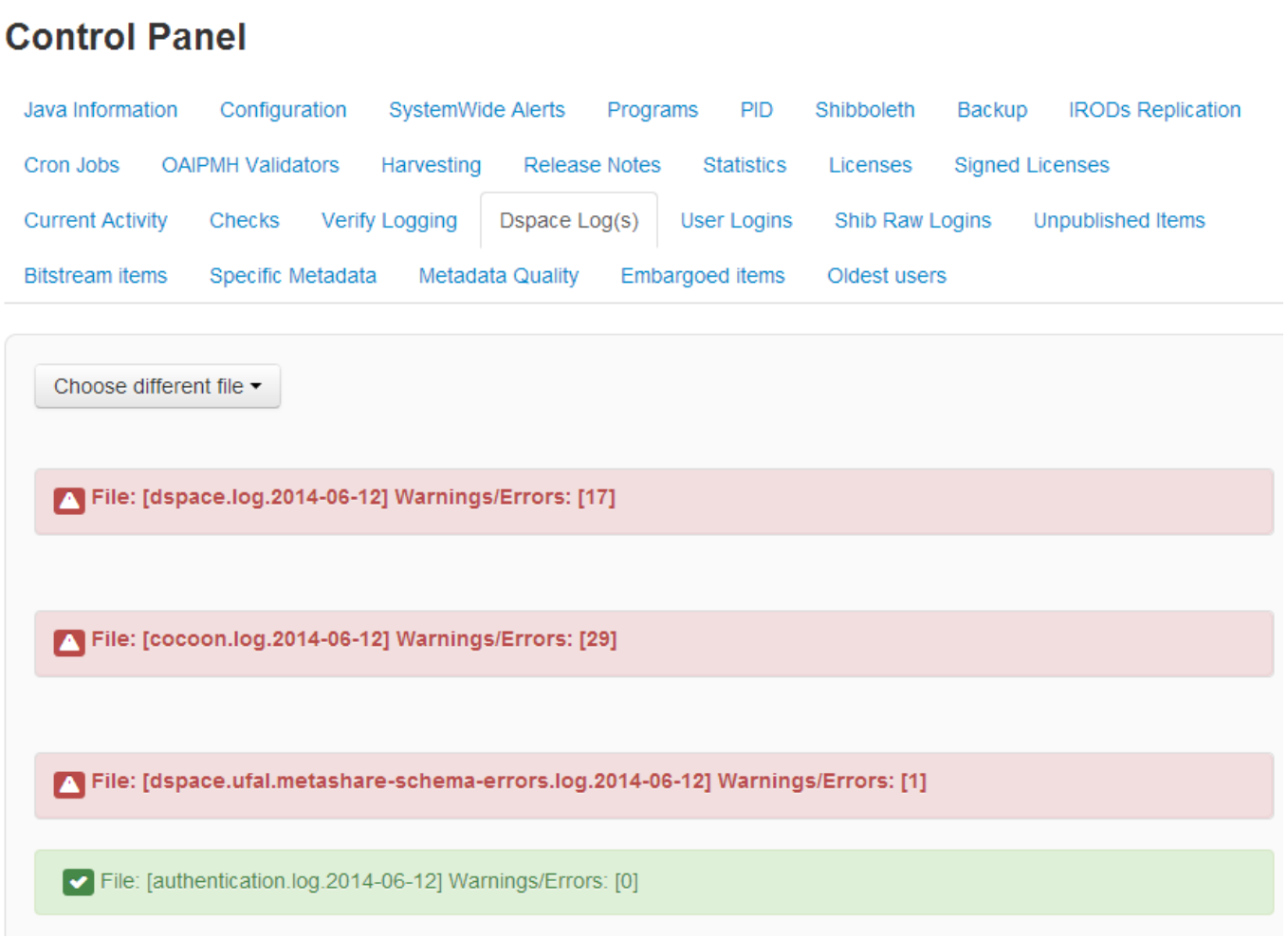
Control Panel - replicas



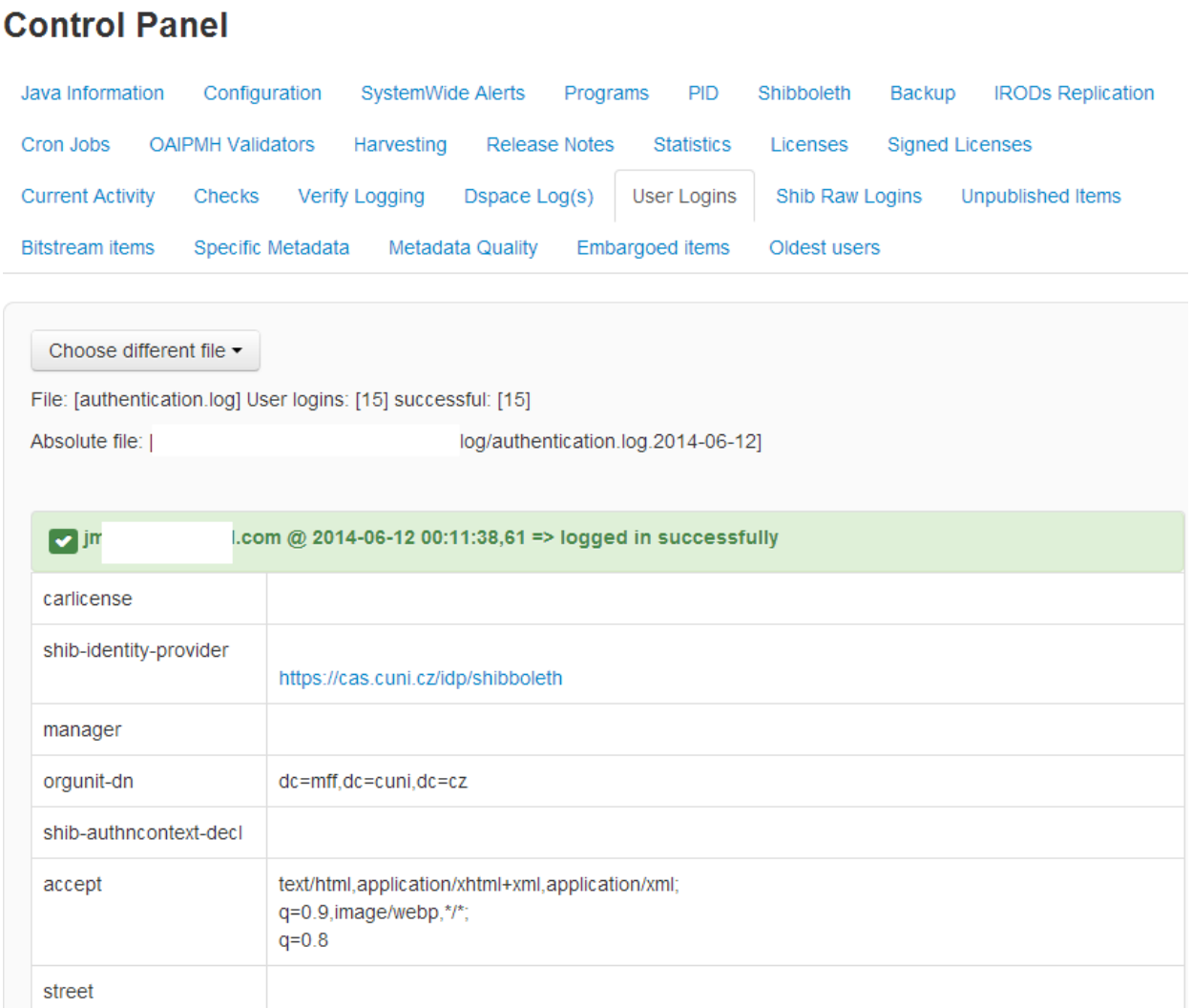
Control Panel - signed licences



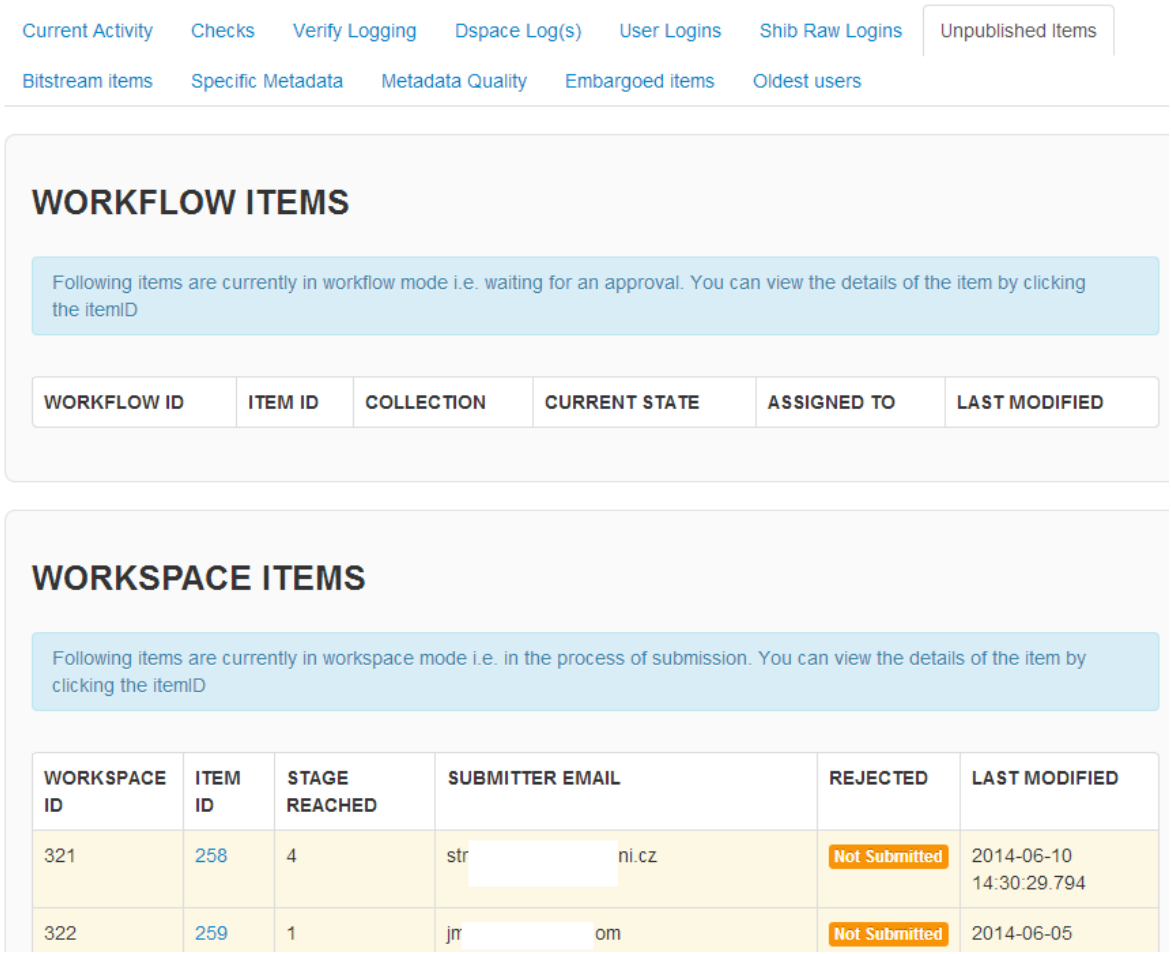
Control Panel - log parsing



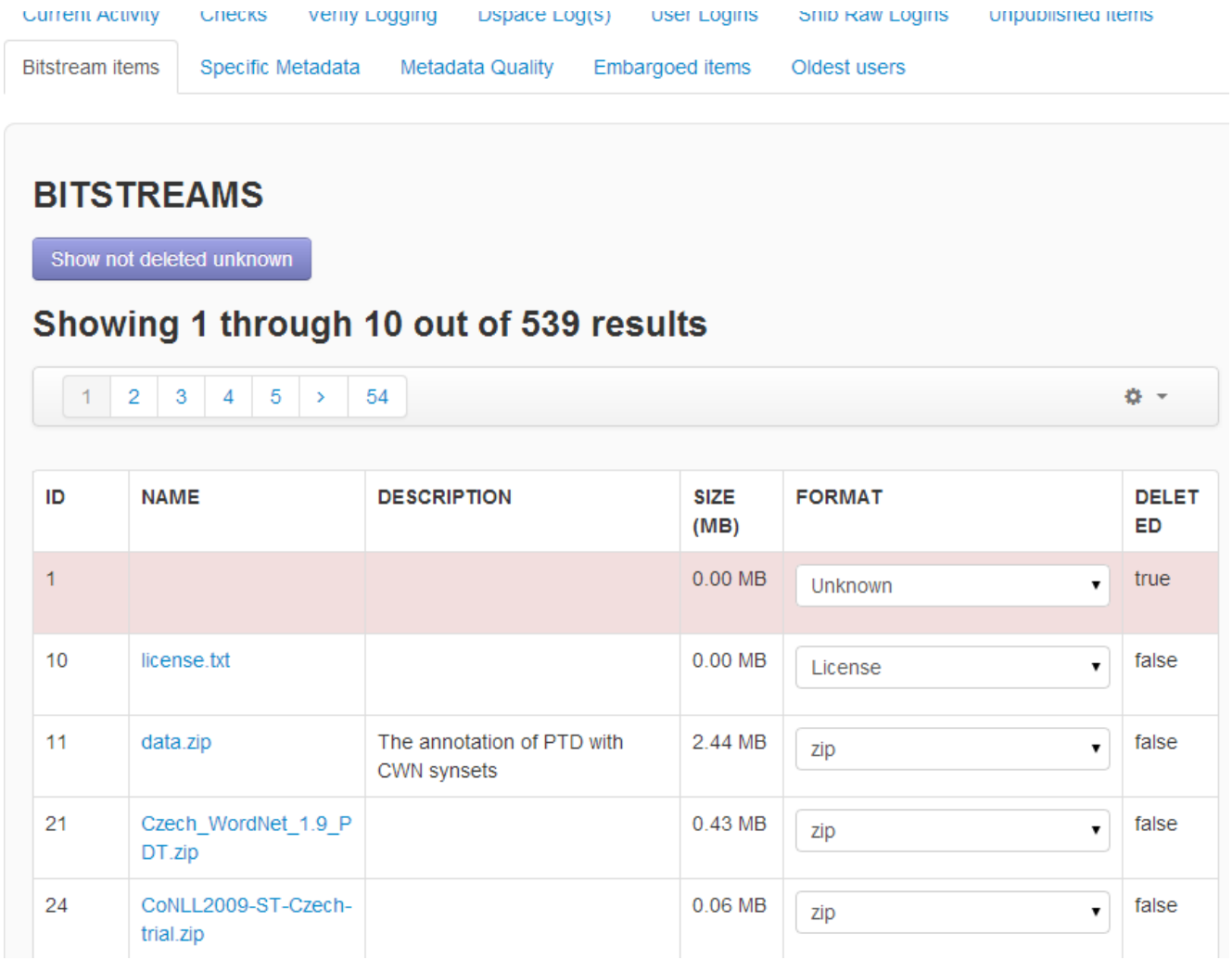
Control Panel - user logins



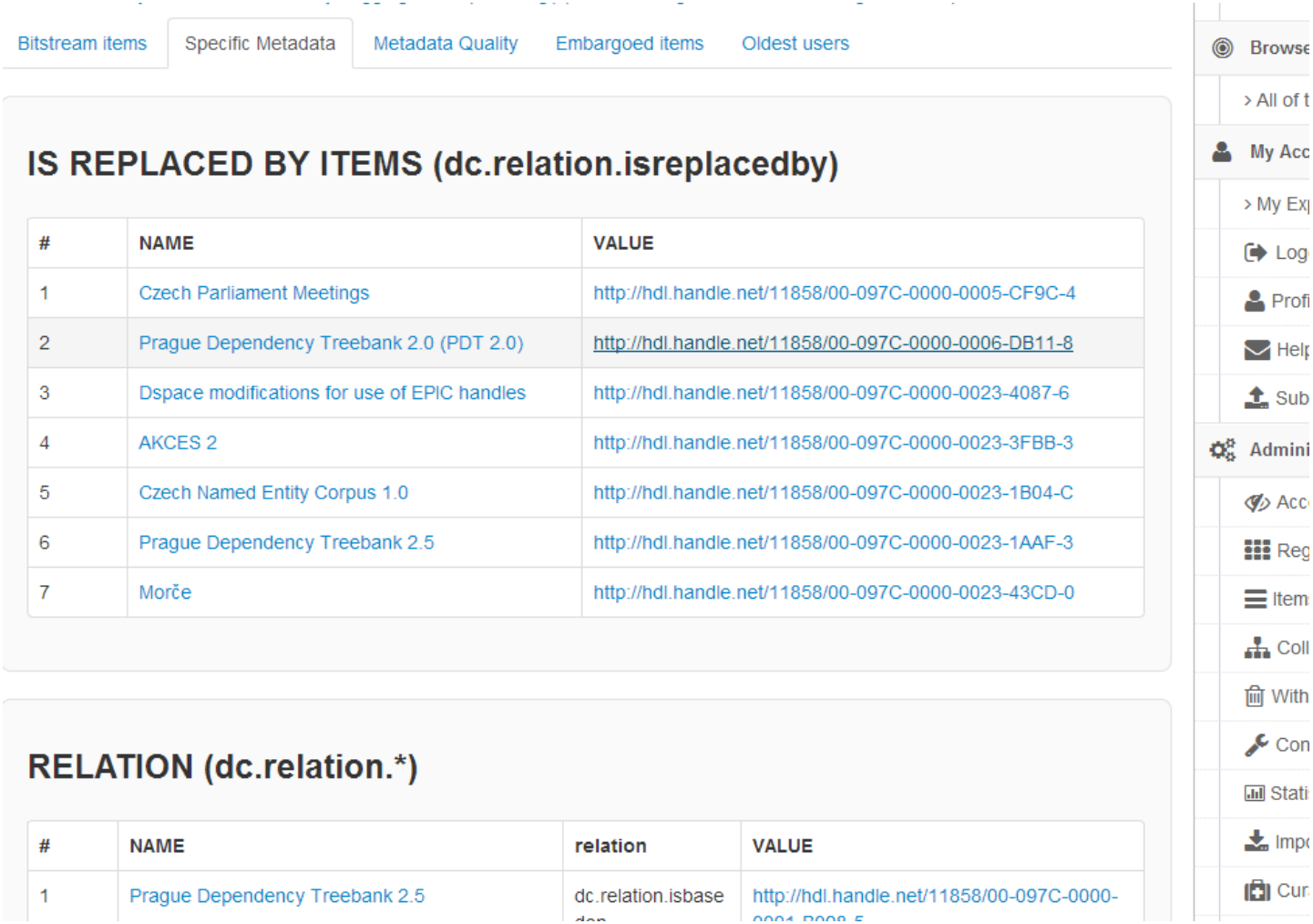
Control Panel - unpublished items



Control Panel - bitstream list



Control Panel - specific metadata



Bitstream items | Specific Metadata | Metadata Quality | Embargoed items | Oldest users

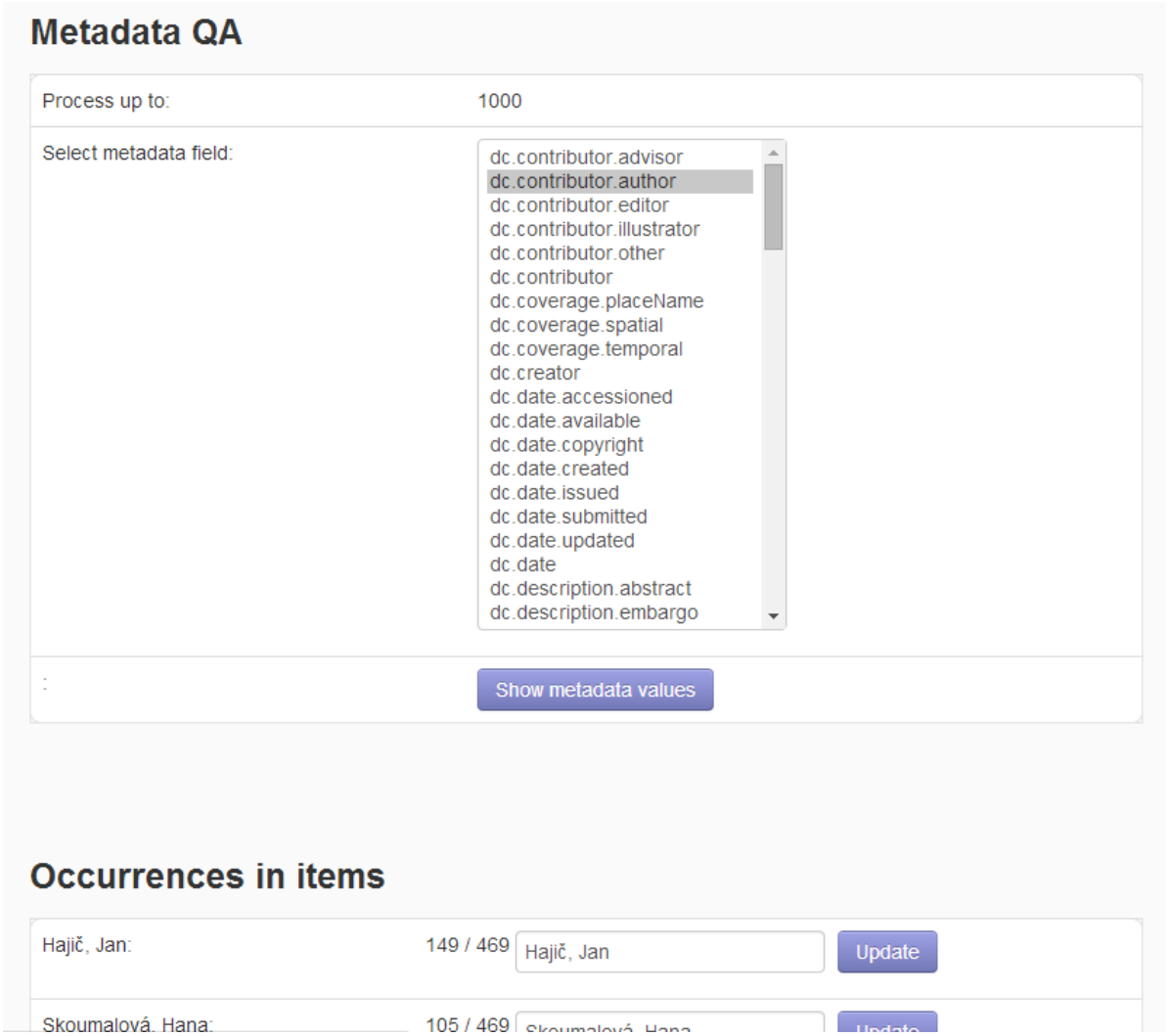
IS REPLACED BY ITEMS (dc.relation.isreplacedby)

#	NAME	VALUE
1	Czech Parliament Meetings	http://hdl.handle.net/11858/00-097C-0000-0005-CF9C-4
2	Prague Dependency Treebank 2.0 (PDT 2.0)	http://hdl.handle.net/11858/00-097C-0000-0006-DB11-8
3	Dspace modifications for use of EPIC handles	http://hdl.handle.net/11858/00-097C-0000-0023-4087-6
4	AKCES 2	http://hdl.handle.net/11858/00-097C-0000-0023-3FBB-3
5	Czech Named Entity Corpus 1.0	http://hdl.handle.net/11858/00-097C-0000-0023-1B04-C
6	Prague Dependency Treebank 2.5	http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3
7	Morče	http://hdl.handle.net/11858/00-097C-0000-0023-43CD-0

RELATION (dc.relation.*)

#	NAME	relation	VALUE
1	Prague Dependency Treebank 2.5	dc.relation.isbase	http://hdl.handle.net/11858/00-097C-0000-

Control Panel - metadata quality



Weekly reports

- cron based email report on the internals and statistics of a DSpace repository including the results of curation tasks and integrity checking
- not only for administrators

https://svn.ms.mff.cuni.cz/redmine/projects/dspace-modifications/repository/revisions/master/entry/sources/dspace-api/src/main/java/cz/cuni/mff/ufal/Report.java

Logger

- crucial objects should use a logger that also sends emails when an exception is logged and not fail silently
  - still have to filter out “insignificant” ones
- trivial but proved very useful
- replace Logger.getLogger with yourLogger.getLogger

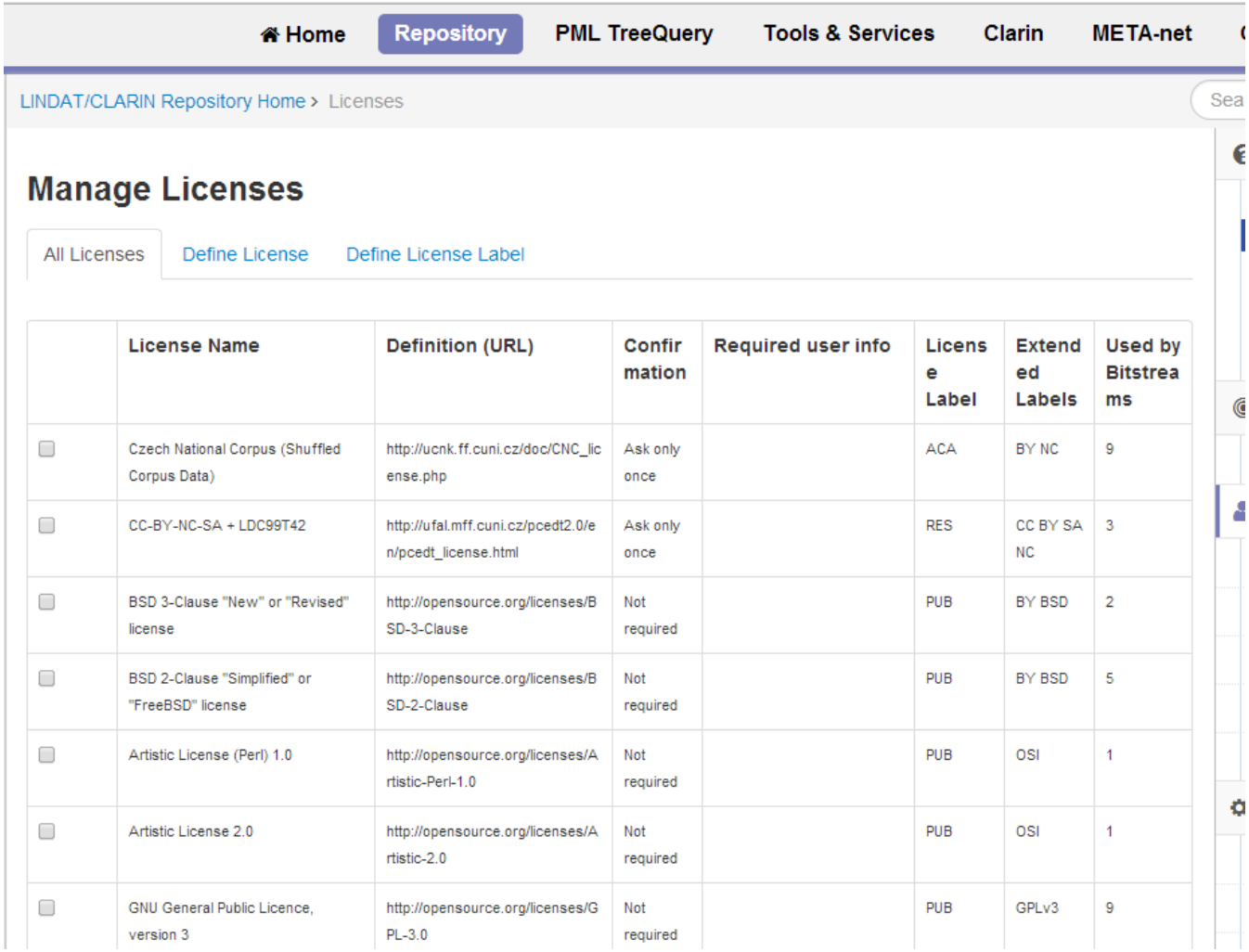
https://svn.ms.mff.cuni.cz/redmine/projects/dspace-modifications/repository/revisions/master/entry/sources/dspace-api/src/main/java/cz/cuni/mff/ufal/Logger.java

Licensing

- create new licences and define restrictions
- attach licences to submissions (bitstreams) in submission workflow
- require authentication and user signature with additional information if necessary
- send download link if necessary

https://svn.ms.mff.cuni.cz/redmine/projects/dspace-modifications/wiki/License_administration

Licensing



Home | Repository | PML TreeQuery | Tools & Services | Clarin | META-net

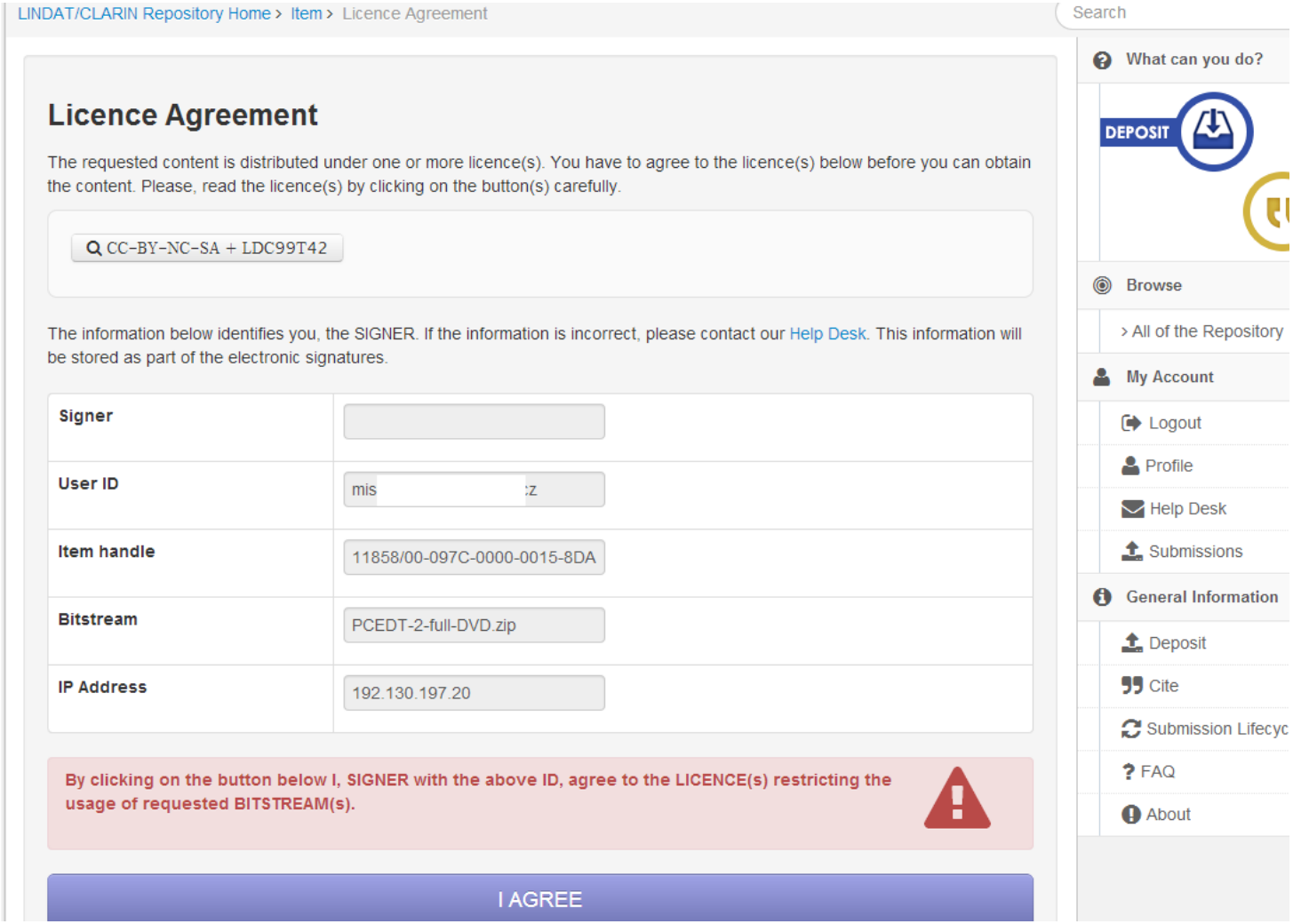
LINDAT/CLARIN Repository Home > Licenses

Manage Licenses

All Licenses | Define License | Define License Label

	License Name	Definition (URL)	Confirmation	Required user info	License Label	Extended Labels	Used by Bitstreams
☐	Czech National Corpus (Shuffled Corpus Data)	http://ucnk.ff.cuni.cz/doc/CNC_license.php	Ask only once		ACA	BY NC	9
☐	CC-BY-NC-SA + LDC99T42	http://ufal.mff.cuni.cz/pcedt2.0/en/pcedt_license.html	Ask only once		RES	CC BY SA NC	3
☐	BSD 3-Clause "New" or "Revised" license	http://opensource.org/licenses/BSD-3-Clause	Not required		PUB	BY BSD	2
☐	BSD 2-Clause "Simplified" or "FreeBSD" license	http://opensource.org/licenses/BSD-2-Clause	Not required		PUB	BY BSD	5
☐	Artistic License (Perl) 1.0	http://opensource.org/licenses/Artistic-Perl-1.0	Not required		PUB	OSI	1
☐	Artistic License 2.0	http://opensource.org/licenses/Artistic-2.0	Not required		PUB	OSI	1
☐	GNU General Public Licence, version 3	http://opensource.org/licenses/GPL-3.0	Not required		PUB	GPLv3	9

Licensing

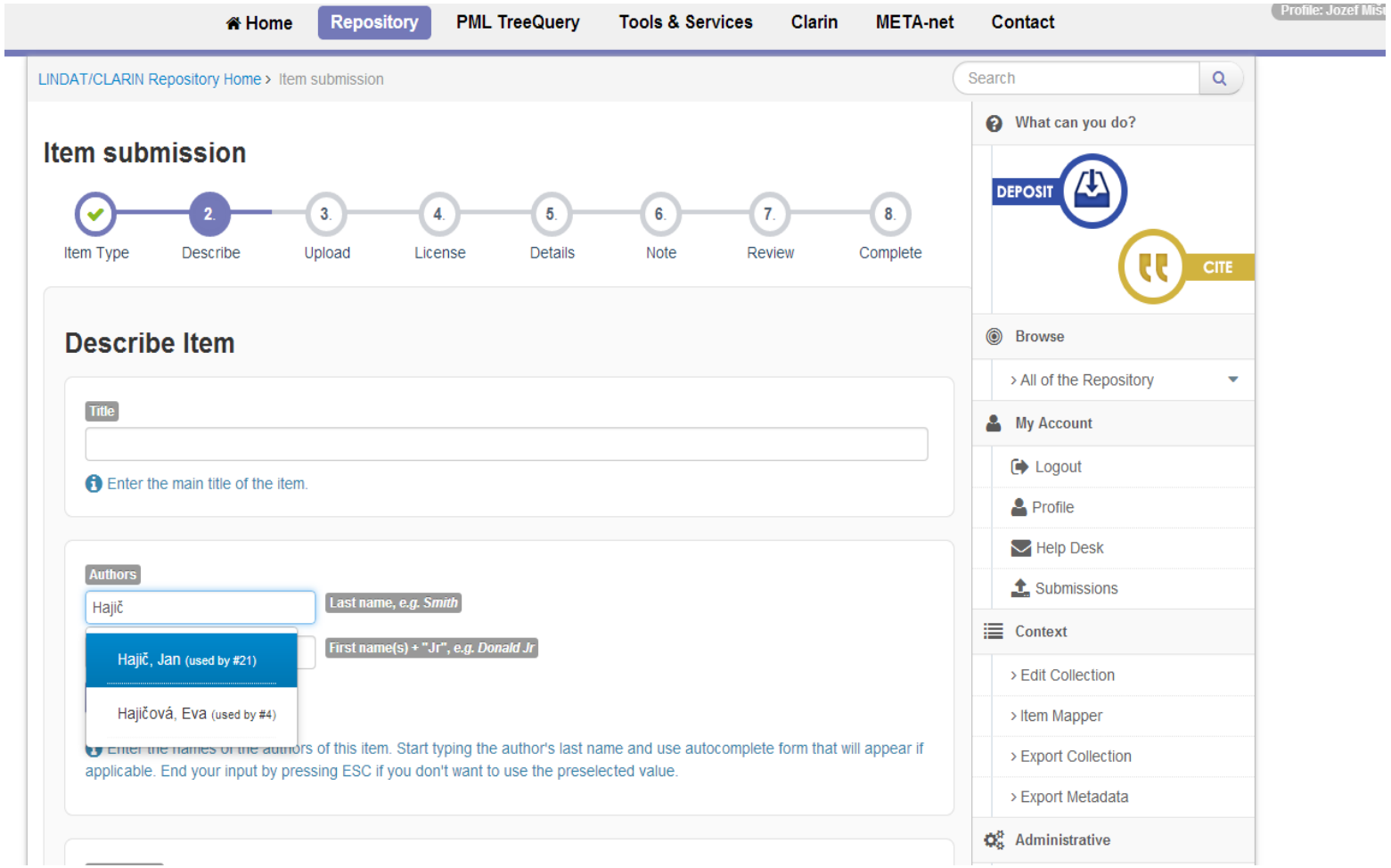

LINDAT/CLARIN Repository Home > Item > Licence Agreement
Search
Licence Agreement
The requested content is distributed under one or more licence(s). You have to agree to the licence(s) below before you can obtain the content. Please, read the licence(s) by clicking on the button(s) carefully.
CC-BY-NC-SA + LDC99T42
The information below identifies you, the SIGNER. If the information is incorrect, please contact our Help Desk. This information will be stored as part of the electronic signatures.
Signer
User ID
mis :z
Item handle
11858/00-097C-0000-0015-8DA
Bitstream
PCEDT-2-full-DVD.zip
IP Address
192.130.197.20
By clicking on the button below I, SIGNER with the above ID, agree to the LICENCE(s) restricting the usage of requested BITSTREAM(s).
I AGREE
What can you do?
DEPOSIT
Browse
> All of the Repository
My Account
Logout
Profile
Help Desk
Submissions
General Information
Deposit
Cite
Submission Lifecyc
FAQ
About

Submission Workflow

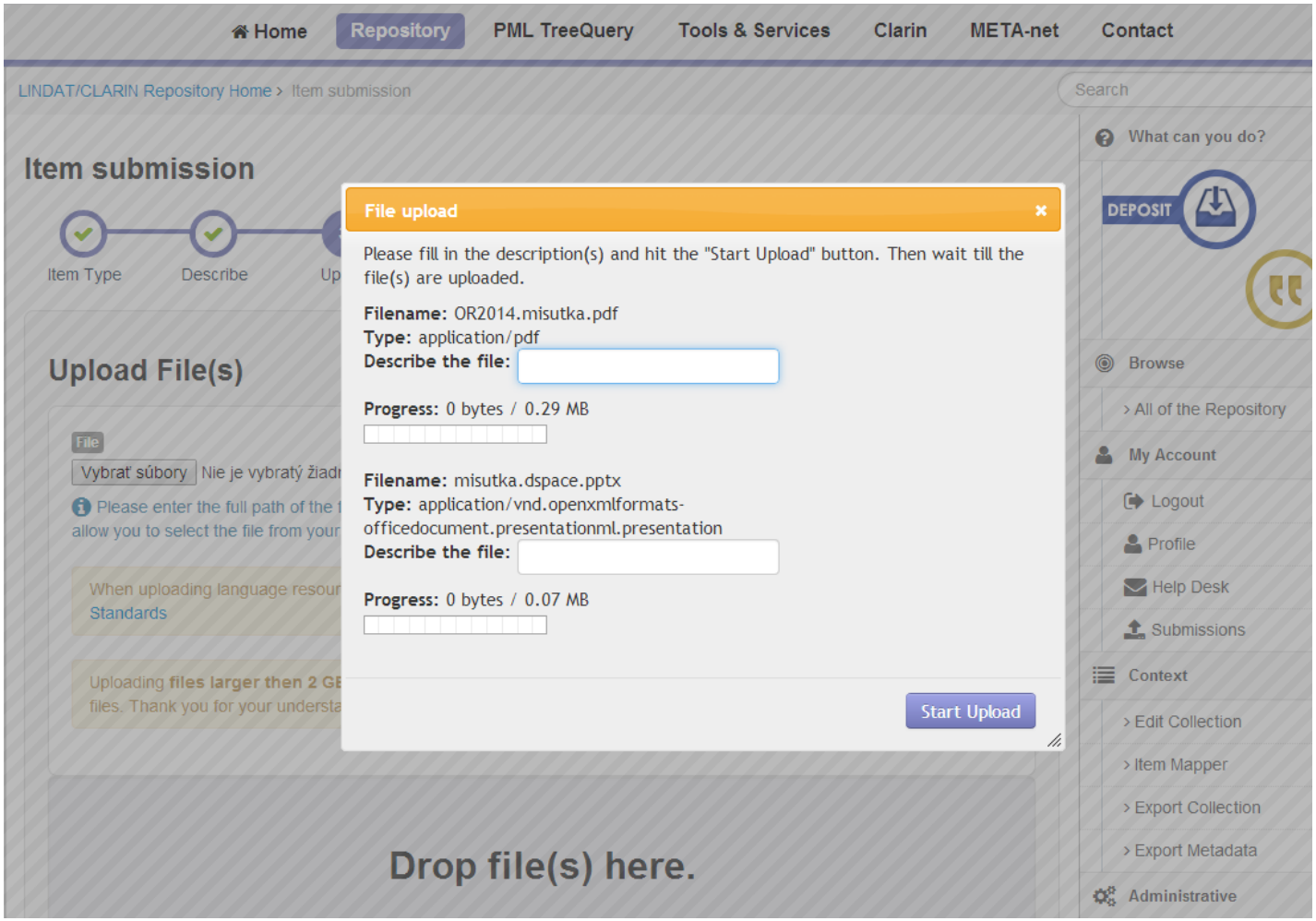
- new UI
- suggest metadata values from other submissions
- autocomplete for vocabularies
- user can leave us a note
- metadata field permission

https://svn.ms.mff.cuni.cz/redmine/projects/dspace-modifications/wiki/FineGrainedFieldPermissions

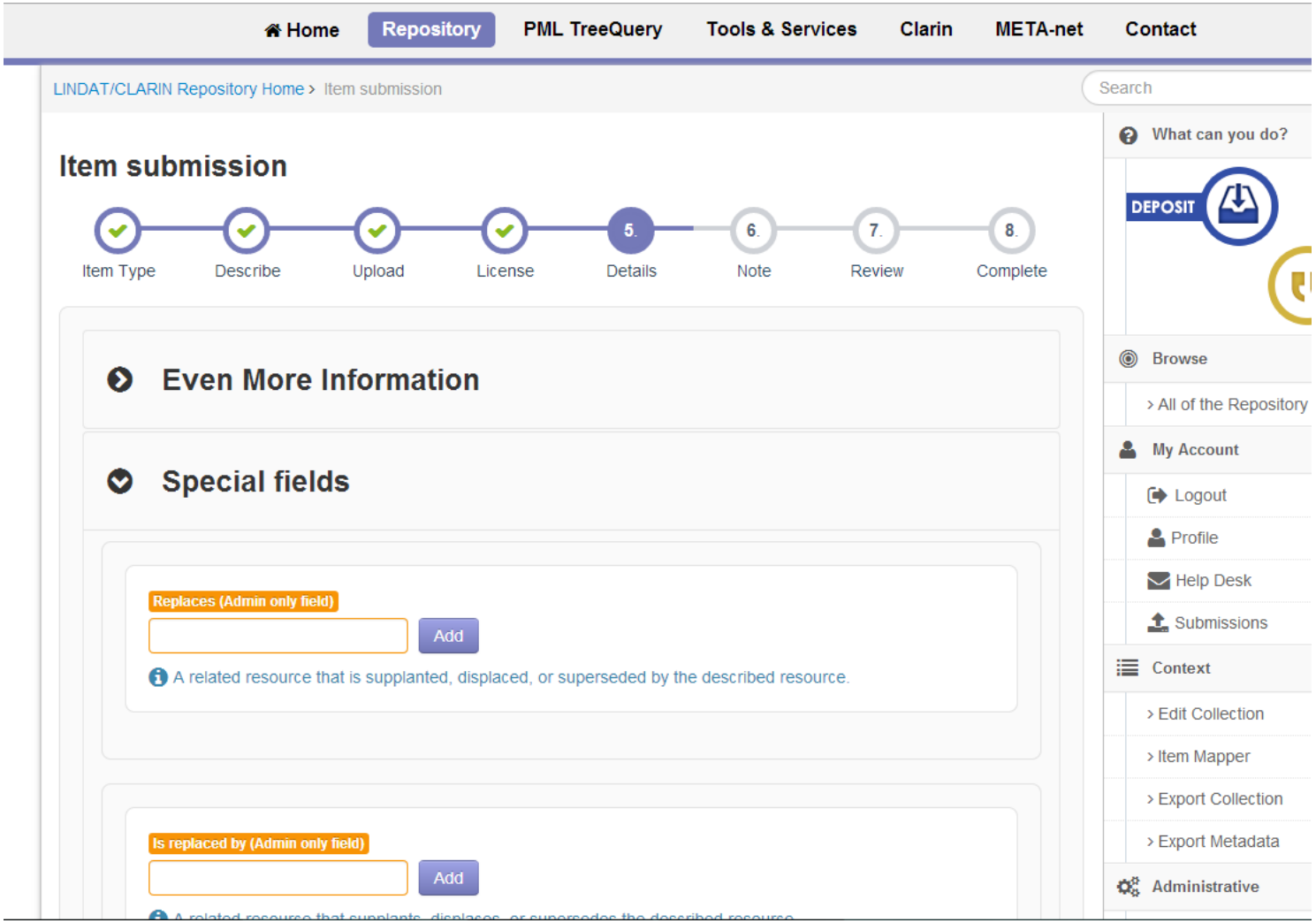
SW - suggest & autocomplete



SW - drag’n’drop



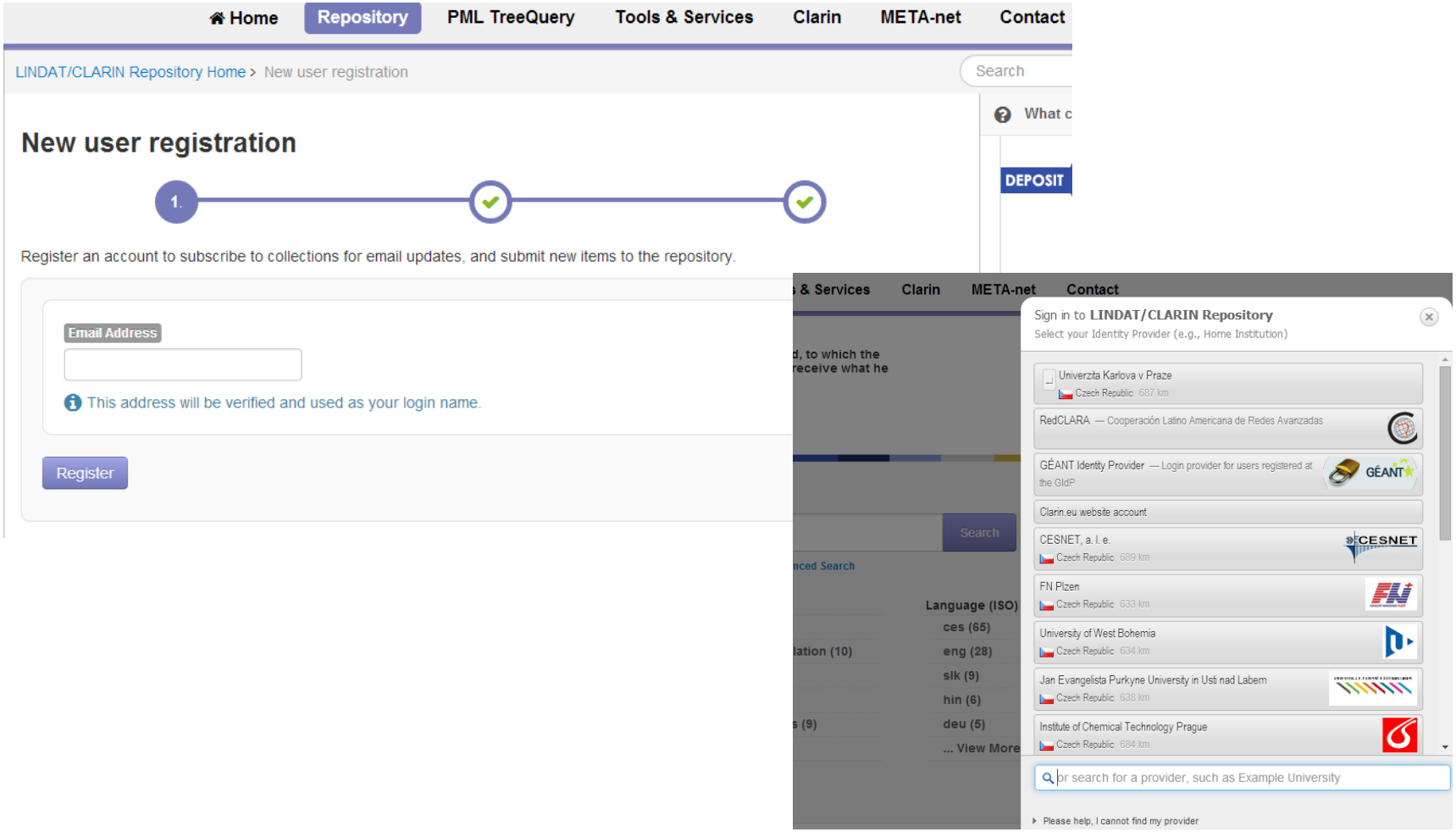
SW - field permission



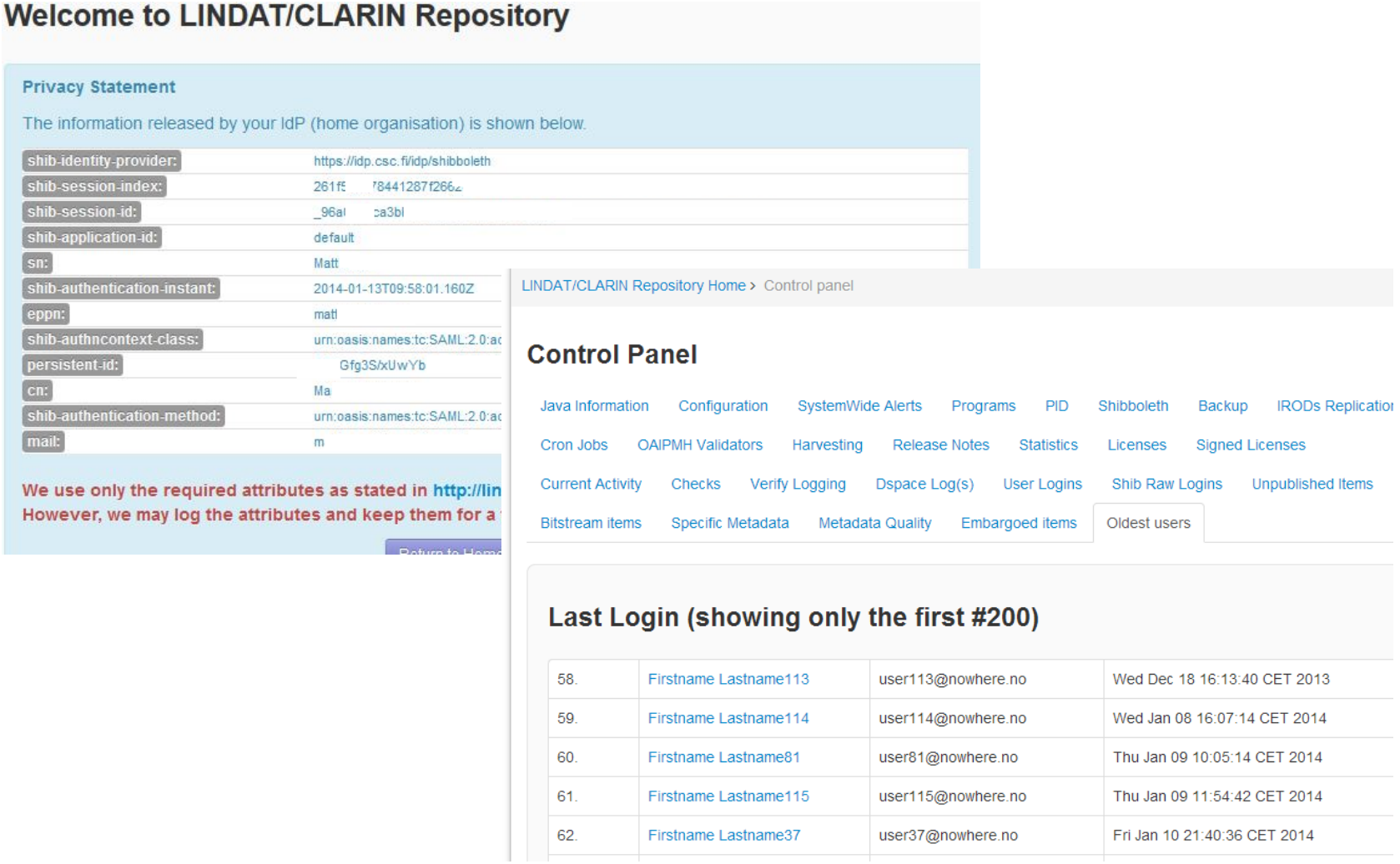
Global AAI

- shibboleth
  - part of NF, CLARIN's SPF, eduGAIN
- local users as fallback mechanism
- reimplementation of shibboleth in DSpace
  - we accept more IDs and **require an email**
- CP integration
- IdP selection in a separate subproject
  - our discovery script used by discojuice

Global AAI


Home
Repository
PML TreeQuery
Tools & Services
Clarin
META-net
Contact
LINDAT/CLARIN Repository Home > New user registration
Search
New user registration
1.
Register an account to subscribe to collections for email updates, and submit new items to the repository.
Email Address
This address will be verified and used as your login name.
Register
DEPOSIT
Sign in to LINDAT/CLARIN Repository
Select your Identity Provider (e.g., Home Institution)
Univerzita Karlova v Praze
Czech Republic 687 km
RedCLARA — Cooperación Latino Americana de Redes Avanzadas
GÉANT Identity Provider — Login provider for users registered at the GIdP
Clarin.eu website account
CESNET, a. l. e.
Czech Republic 689 km
FN Plzen
Czech Republic 633 km
University of West Bohemia
Czech Republic 634 km
Jan Evangelista Purkyne University in Usti nad Labem
Czech Republic 638 km
Institute of Chemical Technology Prague
Czech Republic 684 km
or search for a provider, such as Example University
Please help, I cannot find my provider
Language (ISO)
ces (65)
eng (28)
slk (9)
hin (6)
deu (5)
... View More

Privacy concerns - data protection

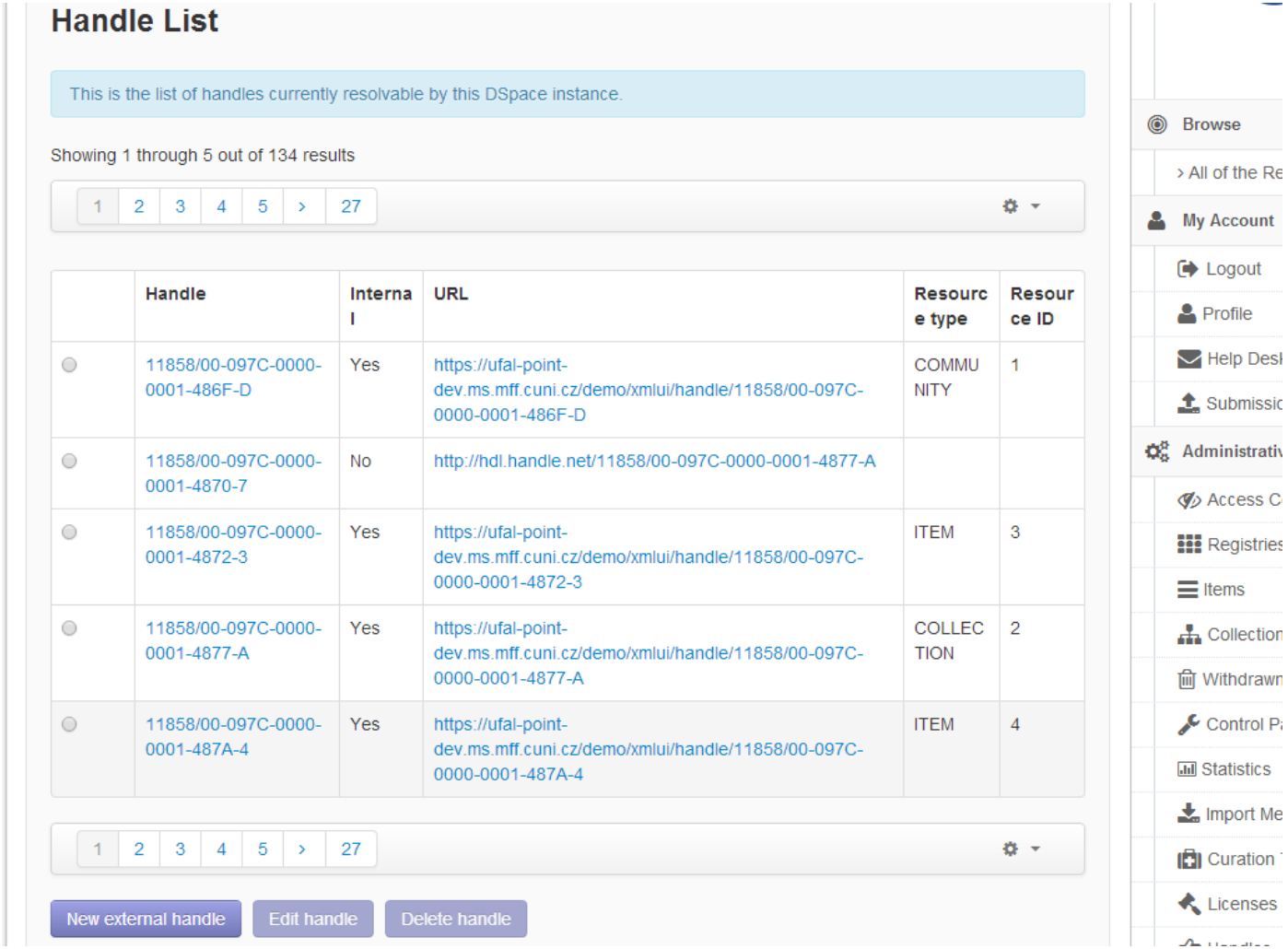


PIDs

- LINDAT/CLARIN uses handles
  - local (more prefixes on one instance)
  - remote through EPIC using REST API
- we support external handles too
  - not connected to DSpace Item but to URL

https://svn.ms.mff.cuni.cz/redmine/projects/dspace-modifications/wiki/New_handle_functionality_in_DSpace

PIDs



Handle List

This is the list of handles currently resolvable by this DSpace instance.

Showing 1 through 5 out of 134 results

1 2 3 4 5 › 27

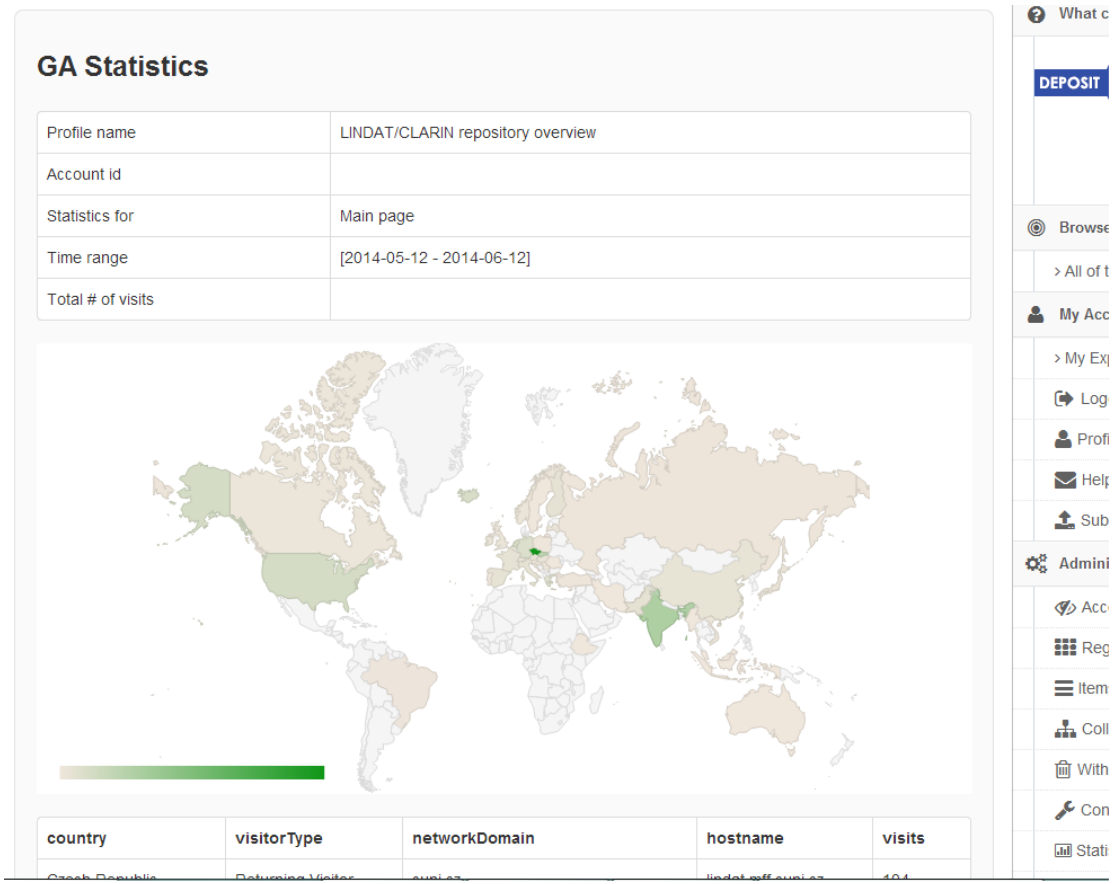
	Handle	Internal	URL	Resource type	Resource ID
○	11858/00-097C-0000-0001-486F-D	Yes	https://ufal-point-dev.ms.mff.cuni.cz/demo/xmlui/handle/11858/00-097C-0000-0001-486F-D	COMMUNITY	1
○	11858/00-097C-0000-0001-4870-7	No	http://hdl.handle.net/11858/00-097C-0000-0001-4877-A		
○	11858/00-097C-0000-0001-4872-3	Yes	https://ufal-point-dev.ms.mff.cuni.cz/demo/xmlui/handle/11858/00-097C-0000-0001-4872-3	ITEM	3
○	11858/00-097C-0000-0001-4877-A	Yes	https://ufal-point-dev.ms.mff.cuni.cz/demo/xmlui/handle/11858/00-097C-0000-0001-4877-A	COLLECTION	2
○	11858/00-097C-0000-0001-487A-4	Yes	https://ufal-point-dev.ms.mff.cuni.cz/demo/xmlui/handle/11858/00-097C-0000-0001-487A-4	ITEM	4

1 2 3 4 5 › 27

New external handle | Edit handle | Delete handle

Browse
› All of the Re
My Account
Logout
Profile
Help Desk
Submissio
Administrativ
Access C
Registries
Items
Collection
Withdrawn
Control P
Statistics
Import Me
Curation
Licenses

GA Statistics

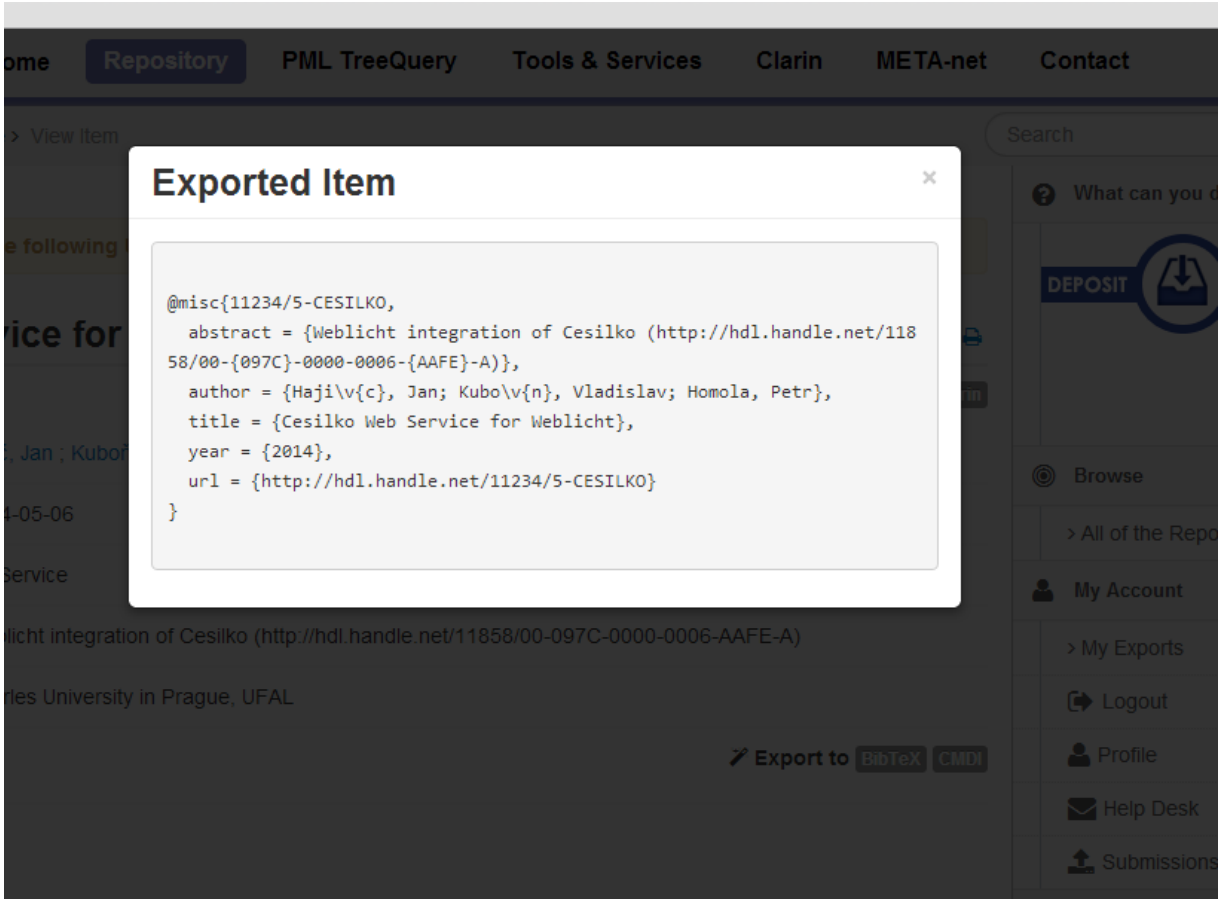


https://svn.ms.mff.cuni.cz/redmine/projects/dspace-modifications/repository/changes/sources/dspace-xmlui/dspace-xmlui-api/src/main/java/cz/cuni/mff/ufal/dspace/app/xmlui/aspect/statistics/GAStatisticsTransformer.java?rev=master

Metadata and dissemination

- basic support of hierarchical metadata - CMDI
- use of local.* metadata not shown to user
- provenance for every change
- exporting to CMDI/bibtex using stripped version of oai processor
  - available at dspace.url/oaistripped

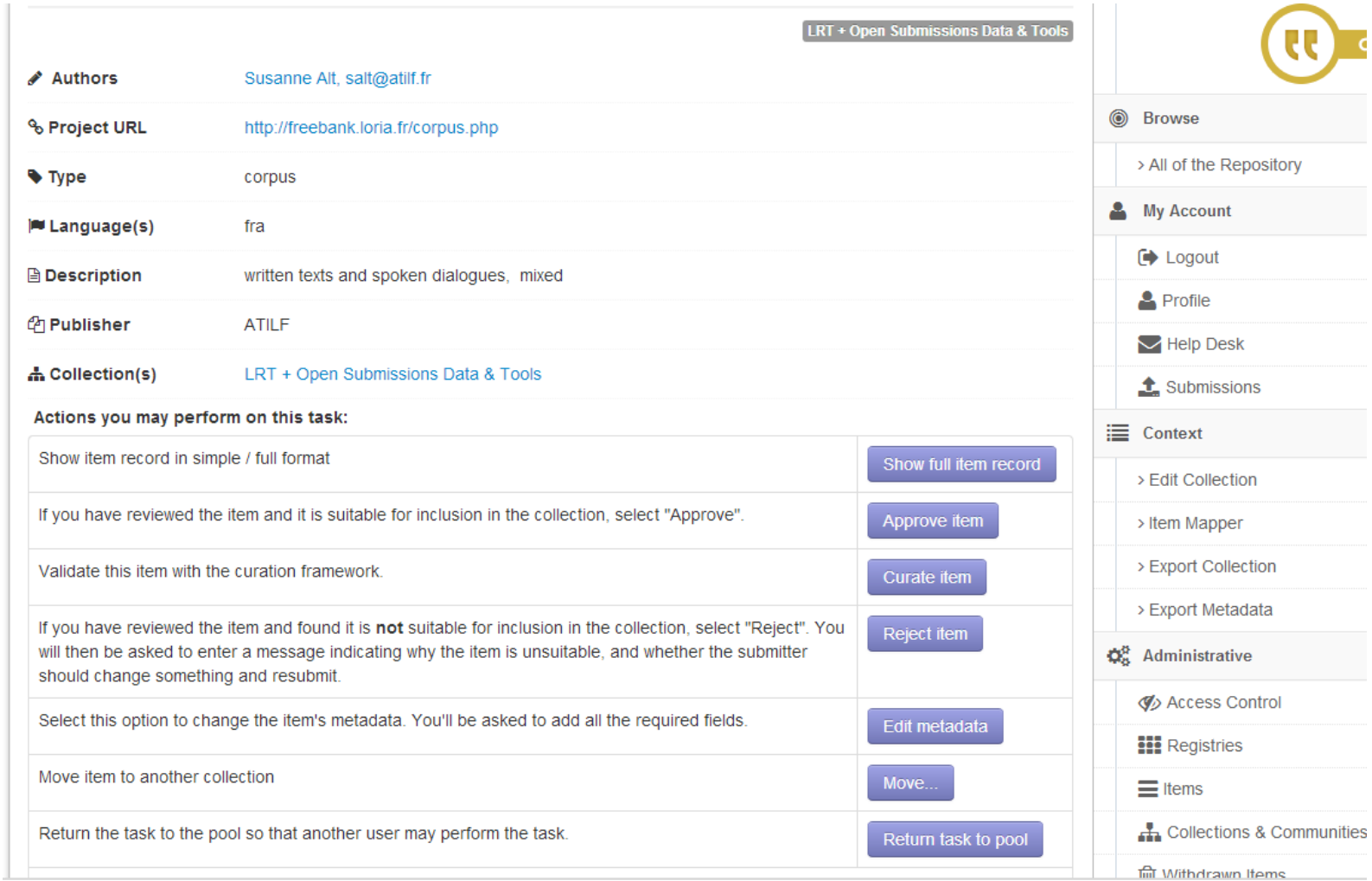
Export to bibtex in Item View



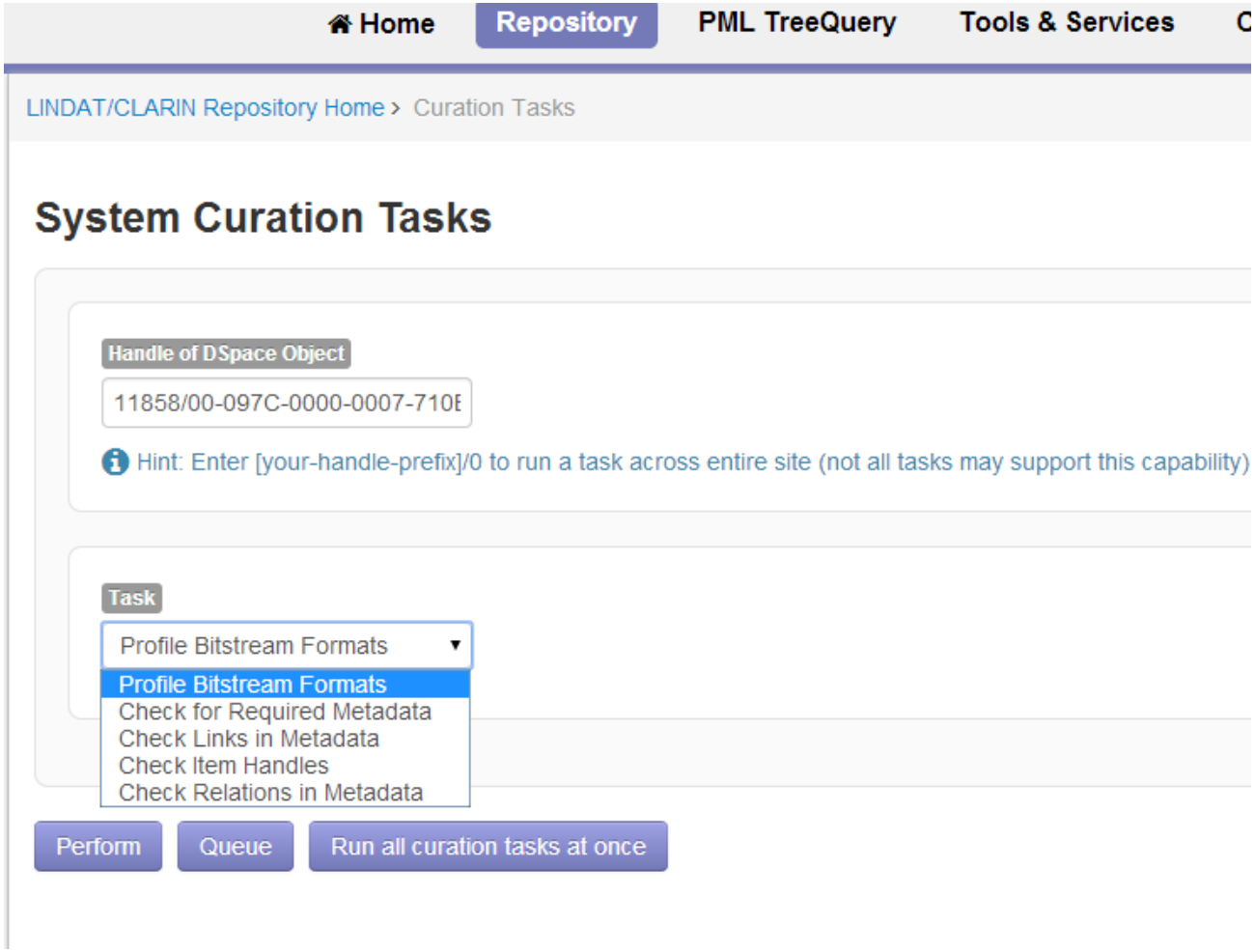
Curation

- allow curation on Workflow objects
- run curation from CP
  - updated error reporting

Curation


LRT + Open Submissions Data & Tools
Authors
Susanne Alt, salt@atilf.fr
Project URL
http://freebank.loria.fr/corpus.php
Type
corpus
Language(s)
fra
Description
written texts and spoken dialogues, mixed
Publisher
ATILF
Collection(s)
LRT + Open Submissions Data & Tools
Actions you may perform on this task:
Show item record in simple / full format
Show full item record
If you have reviewed the item and it is suitable for inclusion in the collection, select "Approve".
Approve item
Validate this item with the curation framework.
Curate item
If you have reviewed the item and found it is not suitable for inclusion in the collection, select "Reject". You will then be asked to enter a message indicating why the item is unsuitable, and whether the submitter should change something and resubmit.
Reject item
Select this option to change the item's metadata. You'll be asked to add all the required fields.
Edit metadata
Move item to another collection
Move...
Return the task to the pool so that another user may perform the task.
Return task to pool
Browse
› All of the Repository
My Account
Logout
Profile
Help Desk
Submissions
Context
› Edit Collection
› Item Mapper
› Export Collection
› Export Metadata
Administrative
Access Control
Registries
Items
Collections & Communities

Curation

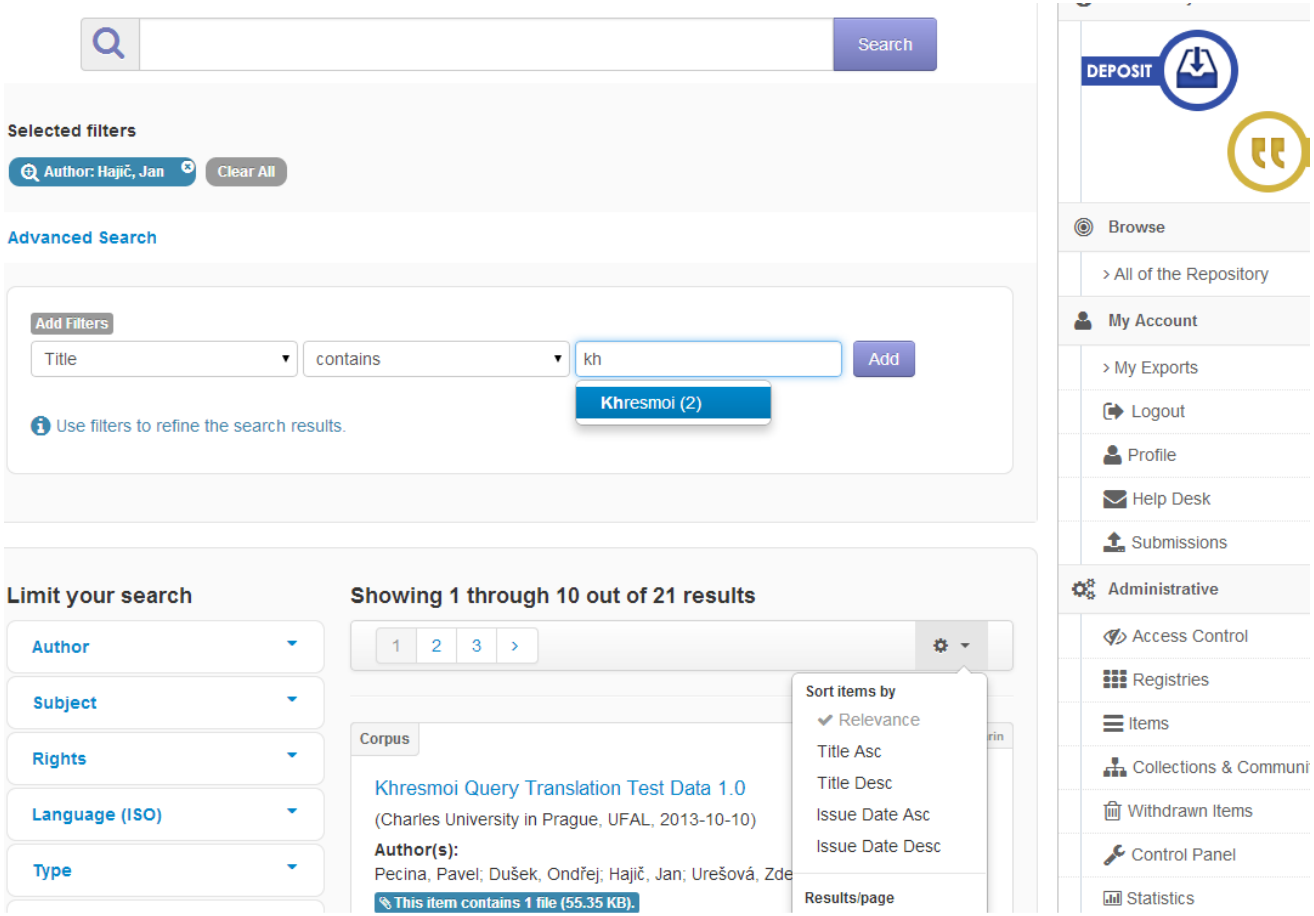

Home
Repository
PML TreeQuery
Tools & Services
LINDAT/CLARIN Repository Home > Curation Tasks
System Curation Tasks
Handle of DSpace Object
11858/00-097C-0000-0007-710E
Hint: Enter [your-handle-prefix]/0 to run a task across entire site (not all tasks may support this capability)
Task
Profile Bitstream Formats
Profile Bitstream Formats
Check for Required Metadata
Check Links in Metadata
Check Item Handles
Check Relations in Metadata
Perform
Queue
Run all curation tasks at once

Searching

- boost matches in title and sort according to relevancy
  - hack but really important (proxy to SOLR)
- add different facets
  - use of local.* metadata automatically generated e.g., collection name

https://svn.ms.mff.cuni.cz/redmine/projects/dspace-modifications/repository/revisions/master/entry/sources/dspace-discovery/dspace-discovery-solr/src/main/java/org/dspace/discovery/SolrServiceImpl.java

Searching


Search
Selected filters
Author: Hajič, Jan
Clear All
Advanced Search
Add Filters
Title
contains
kh
Add
Khresmoi (2)
Use filters to refine the search results.
Limit your search
Author
Subject
Rights
Language (ISO)
Type
Showing 1 through 10 out of 21 results
1
2
3
Sort items by
Relevance
Title Asc
Title Desc
Issue Date Asc
Issue Date Desc
Results/page
Corpus
Khresmoi Query Translation Test Data 1.0
(Charles University in Prague, UFAL, 2013-10-10)
Author(s):
Pecina, Pavel; Dušek, Ondřej; Hajič, Jan; Urešová, Zde
This item contains 1 file (55.35 KB).
DEPOSIT
Browse
All of the Repository
My Account
My Exports
Logout
Profile
Help Desk
Submissions
Administrative
Access Control
Registries
Items
Collections & Communi
Withdrawn Items
Control Panel
Statistics

Demo available

- http://hdl.handle.net/11234/6-DEMO
  - demo-admin
  - demo-user
- master-dev branch

Additional information

- contact us: clarin-list @ ufal.mff.cuni.cz
- issue tracker, wiki, repository
  - http://hdl.handle.net/11234/6-PM
- read-only git repository
  - http://hdl.handle.net/11234/6-VCS
- subprojects
  - https://redmine.ms.mff.cuni.cz/projects/lindat-aai
  - in progress - license selector

Thank you