DSpace in CLARIN

Language Resources and Tools repository

Pavel Straňák

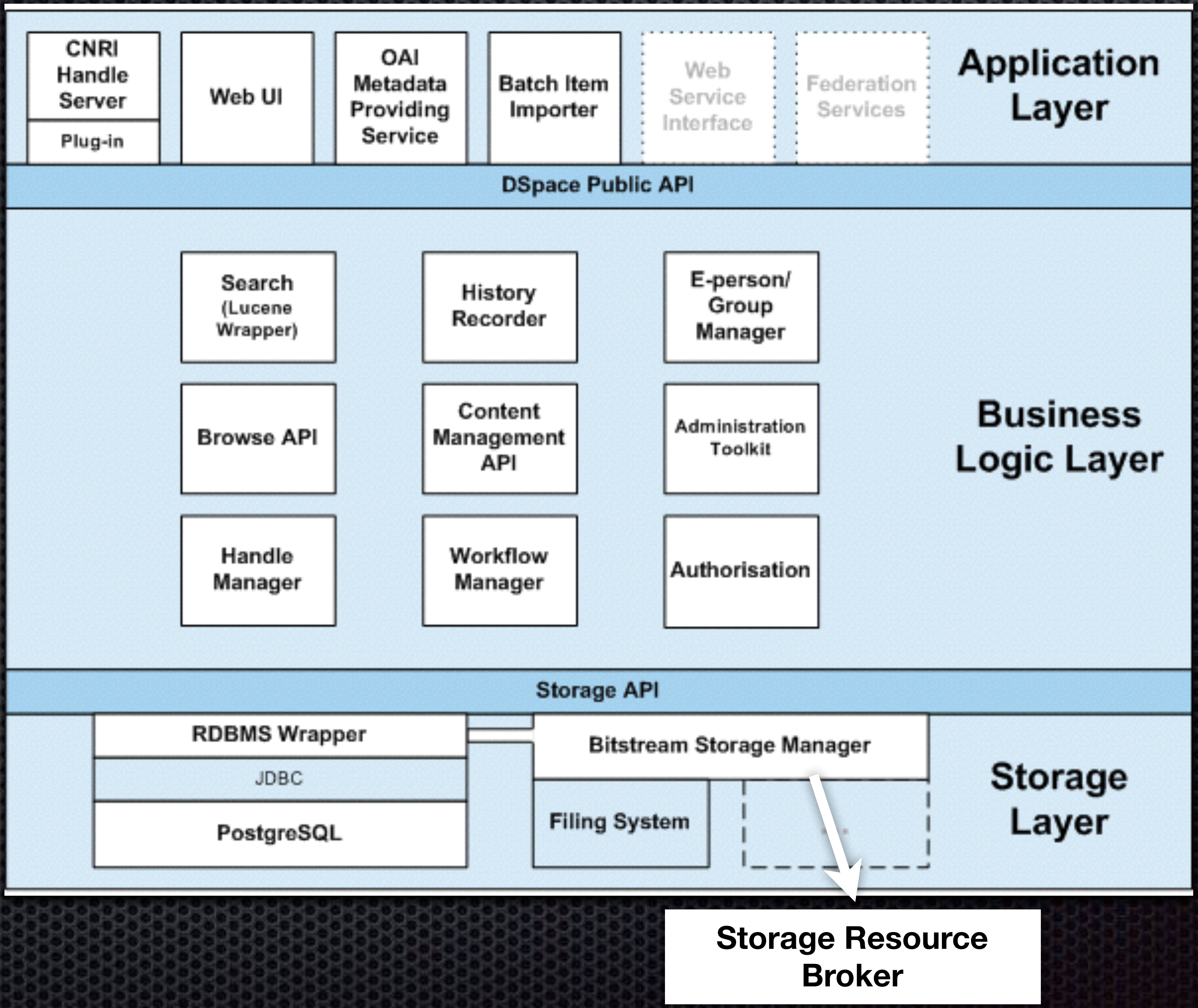
LINDAT/CLARIN

Charles University in Prague

DSpace

- Libraries (incl. national libraries, MIT)
- Image, AV archives
- Museums
- Govt. records, etc.

Dspace architecture



CNRI Handle Server

Plug-in

Web UI

OAI Metadata Providing Service

Batch Item Importer

Web Service Interface

Federation Services

Application Layer

DSpace Public API

Search (Lucene Wrapper)

History Recorder

E-person/ Group Manager

Browse API

Content Management API

Administration Toolkit

Handle Manager

Workflow Manager

Authorisation

Business Logic Layer

Storage API

RDBMS Wrapper

JDBC

PostgreSQL

Bitstream Storage Manager

Filing System

Storage Layer

Storage Resource Broker

Dspace – technical view

- Ready to use “out of the box”
- Free, open source, very customisable
- Very popular (as far as repositories go)
  - > 1000 registered organisations
- Good documentation
- Fair quality of source code (Java)
  - relatively easy to understand and extend

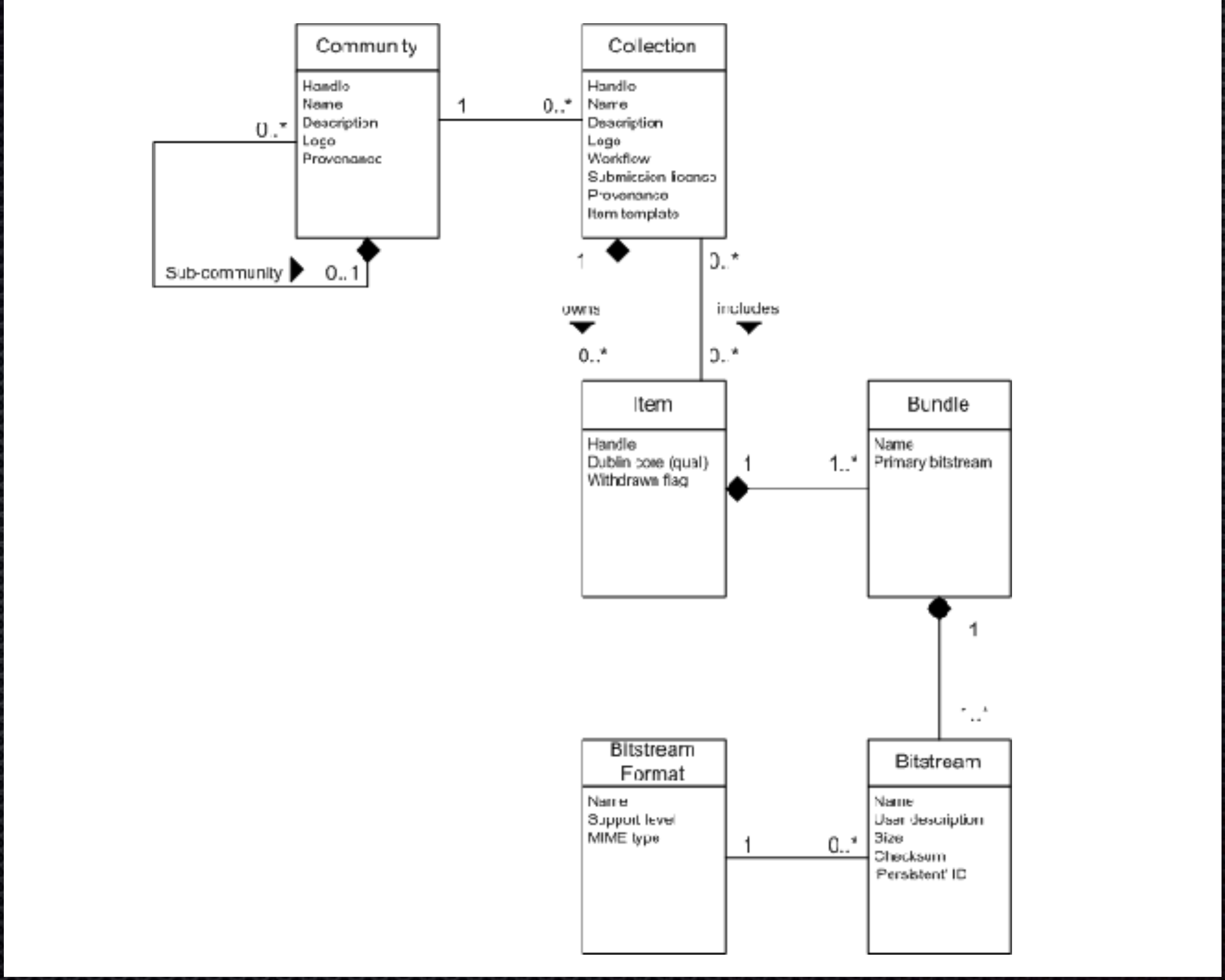
Dspace – users view

- Communities (faculties, projects, data/publications, …)
  - Collections
- Moving records between collections
- One record can appear in multiple collections
- Any data types and formats
  - Supported formats – automated processing possible

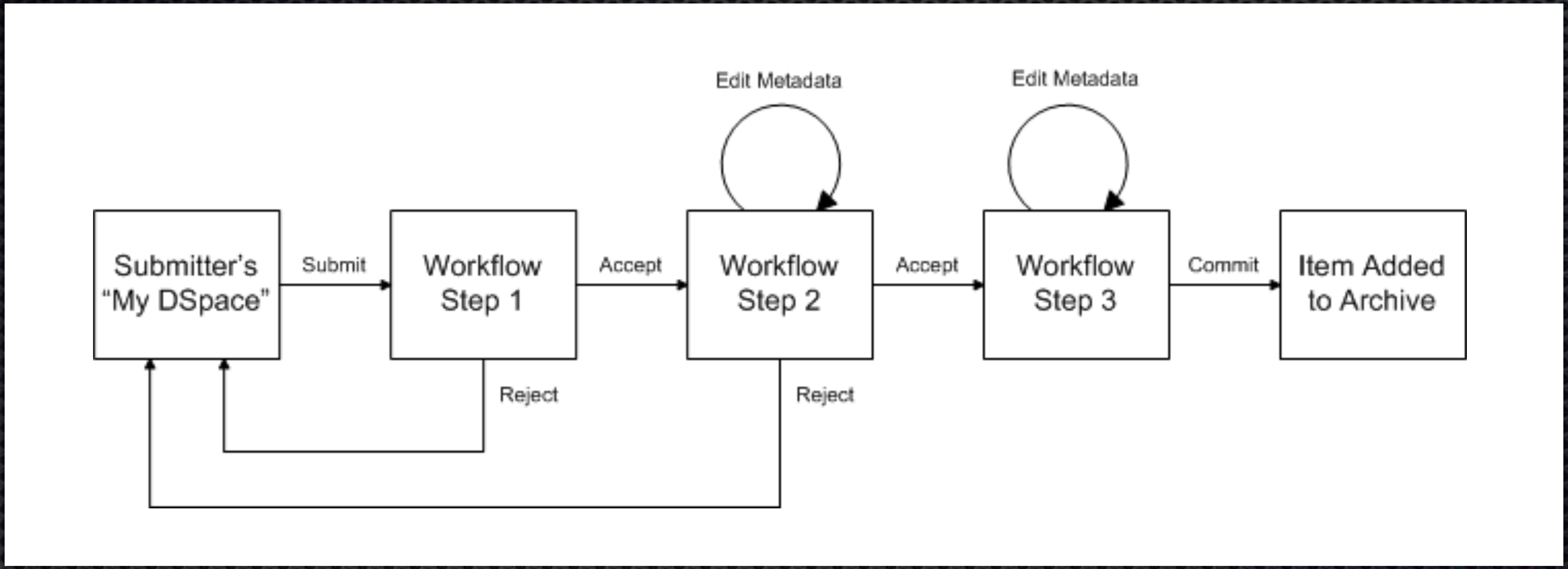
Records, Bitstreams, PIDs

- DSpace includes a Handle server
  - We used EPIC first (adaptation of DSpace), now we use the DSpace handle server
- 1 record = 1 Handle
- Multiple bitstreams per record
  - License, pictures, documentation,
  - Dataset(s)
    - Metadata on Bitstreams (name, description, …)

Dspace Data model



Submission Workflow



DuraSpace

- An umbrella organisation, http://www.duraspace.org/
- Dspace
- FedoraCommons
- DuraCloud
  - online backup, preservation, archiving, streaming, …
- CLARIN cloud? (with EUDAT?)

Important for us in Clarin

- OAI-PMH harvesting compatible
  - includes a harvester too
- Multiple metadata schemata
- Multiple authentication methods (LDAP, SAML2, local accounts, etc.)
  - SAML2 (Shibboleth) used for federated logins

Dspace @ Prague

- Better administration:
  - Control Panel (logins, replication, etc.)
  - License Manager (license signing)
- Hierarchical MD emulation
- Flexible submission workflow
  - pass submission to another user, reserve PID, …

@ Prague – Licenses

- License Manager
- A separate table for “signing”:

user	item	license

- licenses for whole records, per bitstreams not practical
- GUI for creation of licenses
- Attributes for a license:
  - name, URL, symbols like 👤 or $

@ Prague – Submissions

- How to get rich metadata?
  - Minimise the hassle for users
    - What do we REALLY need?
  - Make it nice and fun,
    use good GUI (suggest, drag&drop, etc.)
  - Automate as much as possible
    - Autocomplete for EU Grants from OpenAIRE, etc.

@ Prague – Submissions

- Trying to simplify acquisition of high quality MD
- Language identifier (ISO 639-3)
  - 7679 language codes
  - use AJAX component with auto-completion
  - filter the list to a common subset (639-1 languages)
- Project name and Identifier (OpenAIRE)
- Author (ORCID)?, what else?

Submissions: data upload

- Drag & drop component
- Automatic parsing of common supported data formats
  - Verification of data
  - Extraction of metadata (?)
    - Duration of MP4 AV files
    - Number of tokens and sentences in PML or CoNLL formats

Submission: MD Schemata

- For META-Share (and possibly other project, that require specific MD schema compliance)
- Add a Submission workflow step with their MD schema
  - mapping of MD from previous steps
  - hiding elements filled-in by mapping
  - highlighting what remains to be filled-in