Implementing a funder repository with heterogeneous material and advanced presentation capabilities

Ioanna-Ourania Stathopoulou, Nikos Houssos, Panagiotis Stathopoulos,
Despina Hardouveli, Alexandra Roubani,
Alexandros Soumplis, Chrysostomos Nanakos

National Documentation Centre / National Hellenic Research Foundation, Greece

7th International Conference on Open Repositories (OR2012), Edinburgh, 9-13 July 2012

Agenda

- Aims and scope of the project
- Methodology and approach
- Implementation / DSpace extensions
- Conclusions – future work

Project context

- Hellenic Ministry of Education programmes for education and lifelong learning, co-funded by the EU structural funds
  - Three phases: 1994-1999, 2000-2006, 2007-2013
  - Overall funds in the range of billion Euros
- Activities funded:
  - Face-to-face courses / seminars
  - Creation of educational material
  - Studies, surveys, reports
  - Conferences, workshops
  - Research: theses, collaborative projects

Aims and scope of the project

- Decision by funding authority to create a repository (autumn 2010)
- Aims: Dissemination, reuse and preservation of the digital material produced in the frame of the funded programmes
- Single point of (persistent) access to material, standards-based documentation/metadata, interoperability, link of content to project / funding programme
- Repository available (since spring 2011)
  - http://repository.edulll.gr

Project inputs and outputs

- Input material:
  - Digital files arranged in folders – (hopefully) according to project
  - List of projects
  - No metadata!
- Output aimed for:
  - Repository with standards compliant metadata records along with processed digital files, suitable for dissemination
  - User-friendly interface for human users
  - Programmatic interface (OAI-PMH) for reuse by third-party software applications

Methodology and approach

- Investigate input material
  - Identify content of archival value – exclude all other material (e.g. administrative documents)
  - Assign priorities to content
- Documentation / cataloguing of material according to standards – application profile created
  - Subject cataloguing
  - No self-archiving at this phase
- Digital processing of input files to produce copies suitable for preservation and dissemination – both initial and processed files kept in repository
- A lot of effort allocated to user-friendly interfaces for browsing metadata and accessing content

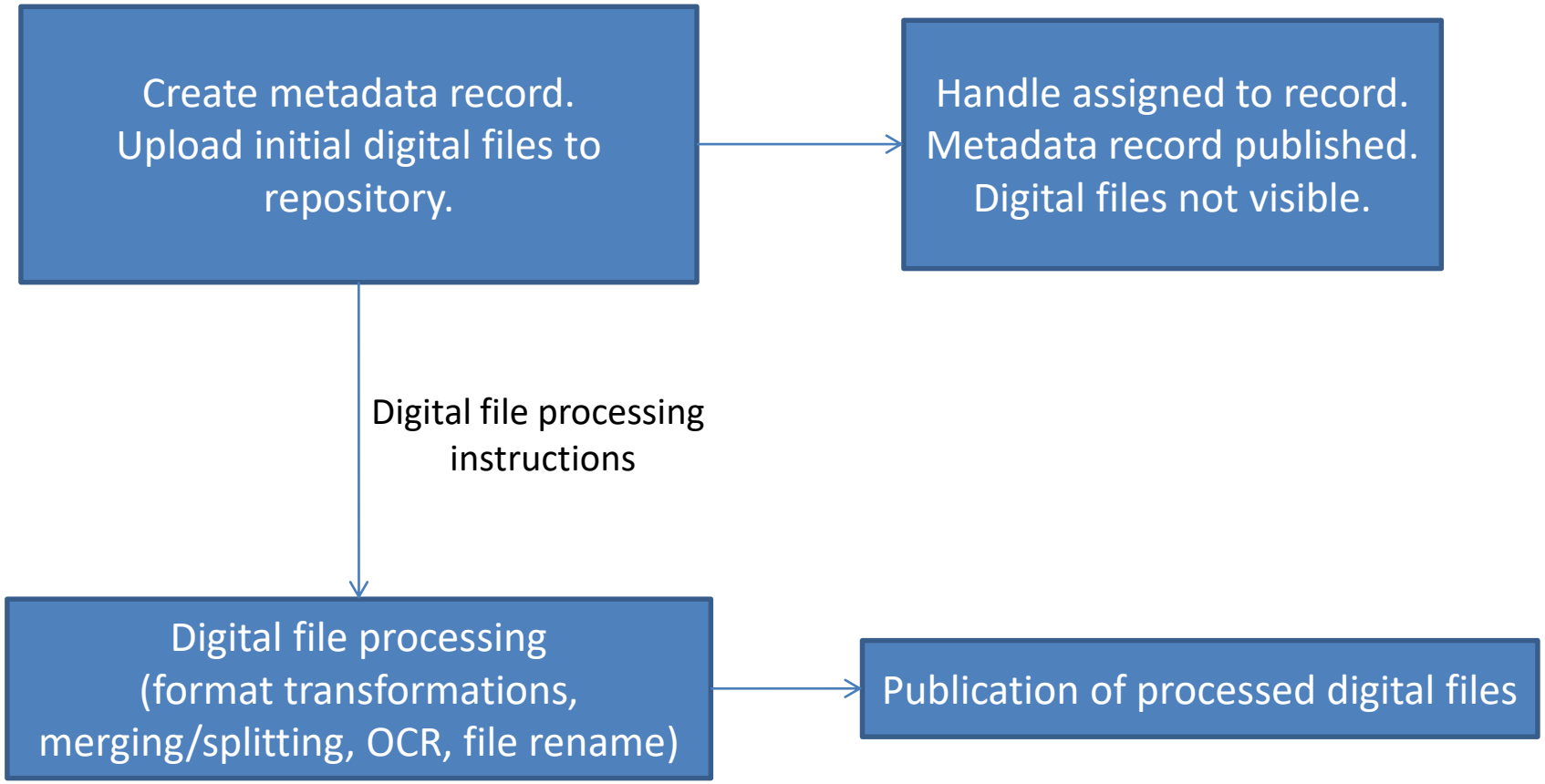
Metadata schema

- Custom application profile developed based on common international schemata
  - Qualified Dublin Core
  - LOM
  - PREMIS
  - CERIF
- Vocabularies used
  - Eurovoc for thematic indexing – due to inter-disciplinarity and multi-linguality
  - Eurydice vocabulary for education levels

Metadata schema

- Custom application profile developed based on common international schemata
  - Qualified Dublin Core
  - LOM
  - PREMIS
  - CERIF
- Vocabularies used
  - Eurovoc for thematic indexing – due to inter-disciplinarity and multi-linguality
  - Eurydice vocabulary for education levels

Processing workflow

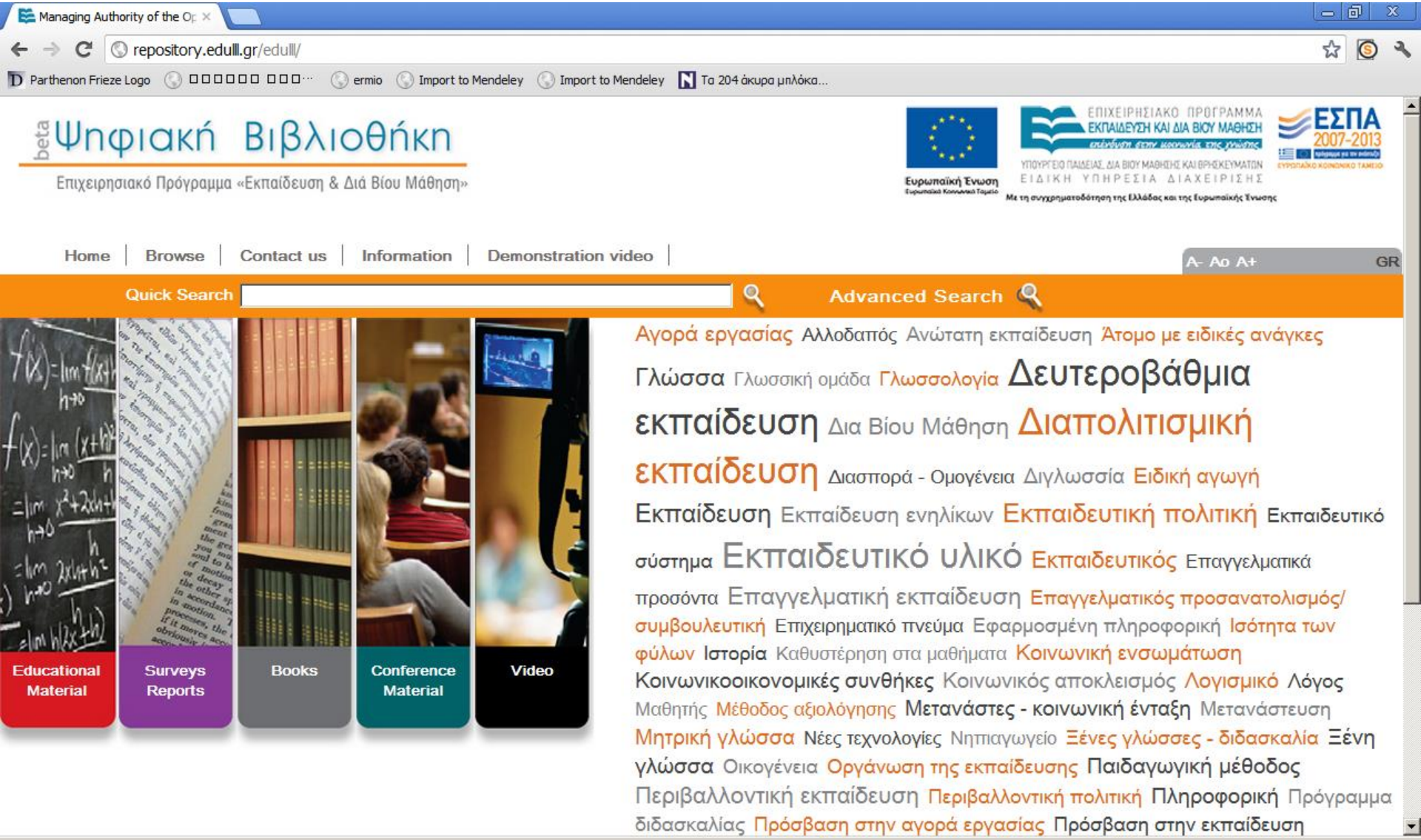


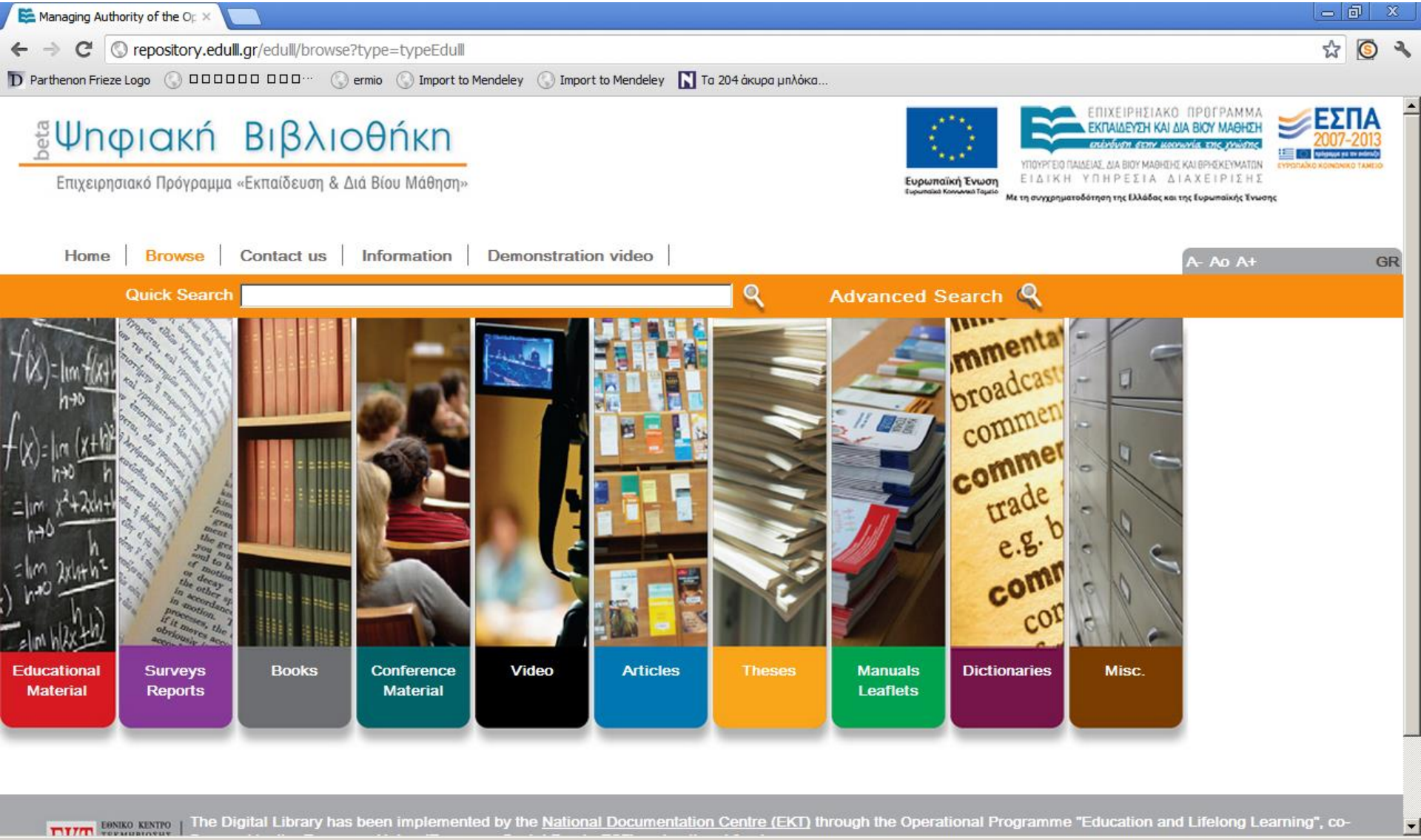
DSpace extensions

- Submission form enhancements
- User-friendly browsing (tagcloud, image-based type browsing)
- Embedding of video files
- Online reading / streaming of text material

Submission form enhancements

- Simplify procedure, fewer steps
- Enter metadata and upload multiple files from single form
- Subsequently edit metadata from submission form
- Multi-lingual controlled vocabularies with user-friendly values
- Various new widgets (e.g. language selection for fields, name processing)


Managing Authority of the Op
repository.edulll.gr/edulll/
Parthenon Frieze Logo
ermio
Import to Mendeley
Import to Mendeley
Τα 204 άκυρα μπλόκα...
beta Ψηφιακή Βιβλιοθήκη
Επιχειρησιακό Πρόγραμμα «Εκπαίδευση & Διά Βίου Μάθηση»
Ευρωπαϊκή Ένωση
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ, ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης
Home
Browse
Contact us
Information
Demonstration video
A- Ao A+
GR
Quick Search
Advanced Search
Educational Material
Surveys Reports
Books
Conference Material
Video
Αγορά εργασίας
Αλλοδαπός
Ανώτατη εκπαίδευση
Άτομο με ειδικές ανάγκες
Γλώσσα
Γλωσσική ομάδα
Γλωσσολογία
Δευτεροβάθμια εκπαίδευση
Δια Βίου Μάθηση
Διαπολιτισμική εκπαίδευση
Διασπορά - Ομογένεια
Διγλωσσία
Ειδική αγωγή
Εκπαίδευση
Εκπαίδευση ενηλίκων
Εκπαιδευτική πολιτική
Εκπαιδευτικό σύστημα
Εκπαιδευτικό υλικό
Εκπαιδευτικός
Επαγγελματικά προσόντα
Επαγγελματική εκπαίδευση
Επαγγελματικός προσανατολισμός/ συμβουλευτική
Επιχειρηματικό πνεύμα
Εφαρμοσμένη πληροφορική
Ισότητα των φύλων
Ιστορία
Καθυστέρηση στα μαθήματα
Κοινωνική ενσωμάτωση
Κοινωνικοοικονομικές συνθήκες
Κοινωνικός αποκλεισμός
Λογισμικό
Λόγος
Μαθητής
Μέθοδος αξιολόγησης
Μετανάστες - κοινωνική ένταξη
Μετανάστευση
Μητρική γλώσσα
Νέες τεχνολογίες
Νηπιαγωγείο
Ξένες γλώσσες - διδασκαλία
Ξένη γλώσσα
Οικογένεια
Οργάνωση της εκπαίδευσης
Παιδαγωγική μέθοδος
Περιβαλλοντική εκπαίδευση
Περιβαλλοντική πολιτική
Πληροφορική
Πρόγραμμα διδασκαλίας
Πρόσβαση στην αγορά εργασίας
Πρόσβαση στην εκπαίδευση


Managing Authority of the Op
repository.edulll.gr/edulll/browse?type=typeEdulll
Parthenon Frieze Logo
ermio
Import to Mendeley
Import to Mendeley
Τα 204 άκυρα μπλόκα...
beta Ψηφιακή Βιβλιοθήκη
Επιχειρησιακό Πρόγραμμα «Εκπαίδευση & Διά Βίου Μάθηση»
Ευρωπαϊκή Ένωση
ΕΠΙΧΕΙΡΗΣΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΕΚΠΑΙΔΕΥΣΗ ΚΑΙ ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗ
ΥΠΟΥΡΓΕΙΟ ΠΑΙΔΕΙΑΣ, ΔΙΑ ΒΙΟΥ ΜΑΘΗΣΗΣ ΚΑΙ ΘΡΗΣΚΕΥΜΑΤΩΝ
ΕΙΔΙΚΗ ΥΠΗΡΕΣΙΑ ΔΙΑΧΕΙΡΙΣΗΣ
ΕΣΠΑ 2007-2013
Με τη συγχρηματοδότηση της Ελλάδας και της Ευρωπαϊκής Ένωσης
Home
Browse
Contact us
Information
Demonstration video
A- Ao A+
GR
Quick Search
Advanced Search
Educational Material
Surveys Reports
Books
Conference Material
Video
Articles
Theses
Manuals Leaflets
Dictionaries
Misc.
The Digital Library has been implemented by the National Documentation Centre (EKT) through the Operational Programme "Education and Lifelong Learning", co-

User-friendly browsing

- Tag cloud for subjects
- Configurable from dspace.dfg – can be used for any other field indexed for browsing
- Image-based browsing for content types

Configurable tag cloud (examples)

```
# Select the browse index to create a tag cloud for in the home page
# Possible values: any of the browse indices
webui.tagcloud.home.bindex = subject
#
# Select the total tags to show
# Possible values: any integer from 1 to infinity
webui.tagcloud.home.maxtags = 50
#
# Should display the score next to each tag?
# Possible values: true | false
webui.tagcloud.home.showscore = false
#
# The score that tags with lower than that will not appear in the rag
cloud
# Possible values: any integer from 1 to infinity
webui.tagcloud.home.cutlevel = 5
```

Embedding video files

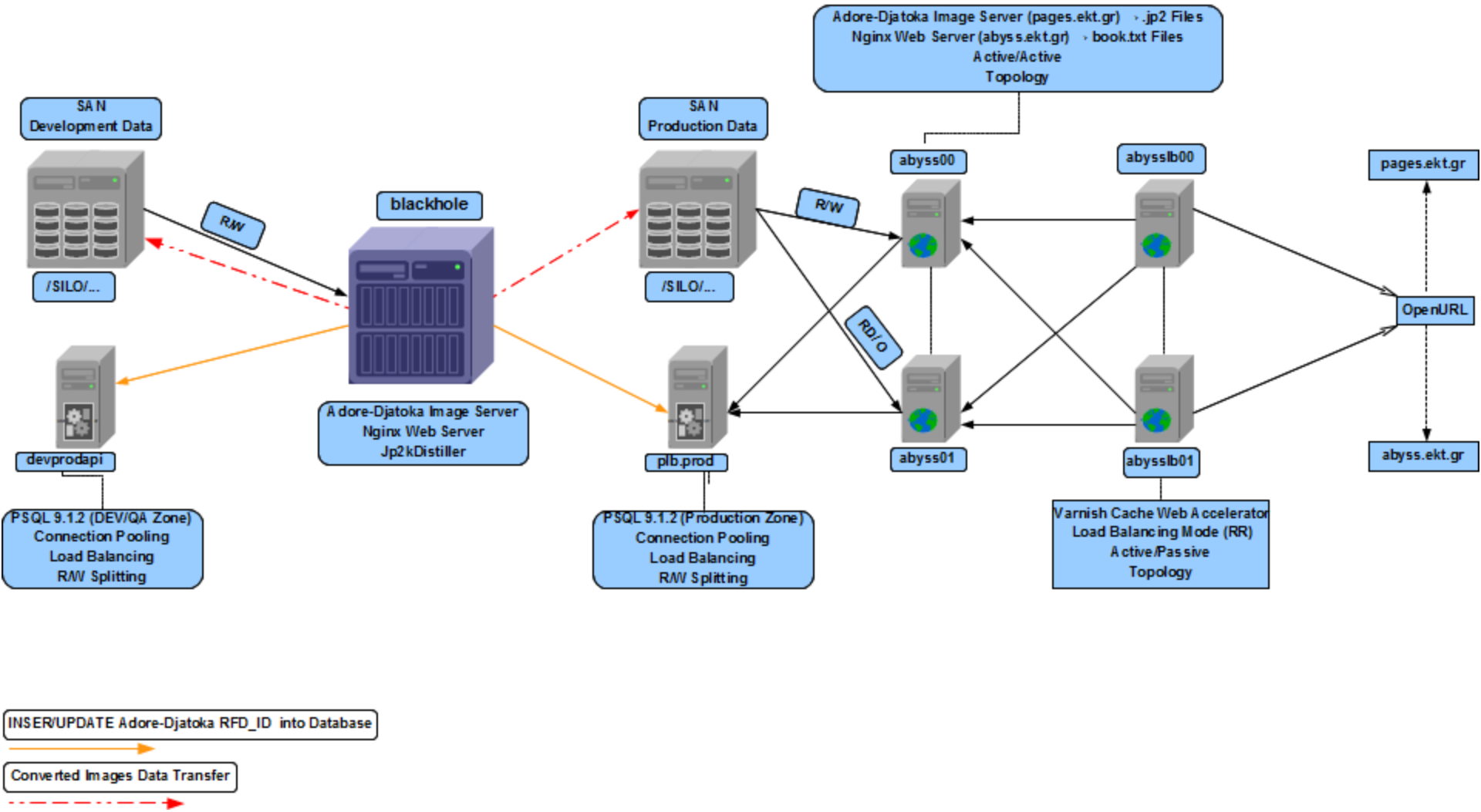
- H.264 enconding using a F4V video file container
- Video stream provided by Woowza
- Flowplayer selected from embedding video on metadata record page on repository
- Video-related fields included to the metadata submission/edit form

Online reader for text

- No need to download big files
- Bookmarking/sharing at the page level
- User access statistics at the page level
- Compatible with multitouch interfaces / tablets
- Based on the Internet Archive Book Reader
- Pages are JPEG2000 files served by an image server
- Sophisticated infrastructure for back-end processing

Online reader backend

- Two-tier backend:
  - content delivery and
  - content transformation
- Content delivery:
  - Nginx/Varnish load balancer/caching to 2 node tomcat/djatoka cluster serving JPEG2000 over a shared SAN storage and utilising a 3 node Postgres 9 cluster
- Content transformation
  - “Shadow” backend with specialized multithreaded content types transformation server
    - After transformation content copied to the delivery backend
  - Interfaced to OJS/Dspace
  - Manages the automatic batch conversion of content
  - Available as OSS (jp2k-distiller) http://code.google.com/p/jp2k-distiller/


Adore-Djatoka Image Server (pages.ekt.gr) › .jp2 Files
Nginx Web Server (abyss.ekt.gr) › book.txt Files
Active/Active
Topology
SAN
Development Data
/SILO/...
R/W
blackhole
Adore-Djatoka Image Server
Nginx Web Server
Jp2kDistiller
devprodapi
PSQL 9.1.2 (DEV/QA Zone)
Connection Pooling
Load Balancing
R/W Splitting
SAN
Production Data
/SILO/...
R/W
RD/O
plb.prod
PSQL 9.1.2 (Production Zone)
Connection Pooling
Load Balancing
R/W Splitting
abyss00
abyss01
abysslb00
abysslb01
Varnish Cache Web Accelerator
Load Balancing Mode (RR)
Active/Passive
Topology
pages.ekt.gr
OpenURL
abyss.ekt.gr
INSER/UPDATE Adore-Djatoka RFD_ID into Database
Converted Images Data Transfer

Jp2k-distiller & dpool – features

- Available as OSS (http://code.google.com/p/dpool/)
- Set of python scripts that schedules and manages batch conversion of content files to different formats, exploiting multithreaded batch conversion features.
  - Need to run multiple encoders/per server in order to achieve high sustainable conversion ratesortunity window lasts until the batch ends.
- Three (3) different conversion options over various Queues
  - PDF/TIFF → PNG Conversion Queue
  - PNG → JP2000 Conversion Queue
  - JP2000 → Publishing Queue
- Interfaced with OJS and DSpace over a HTTP interface
- New version: **dpool** (so radically improved we had to change names! ☺)
  - Tool for modular batch handling of any type of content conversion
  - Truly multinode: a central dispatcher server and an artibitaty amount of distributed "slave" workers over multiple physical or virtual servers
  - Technologies: Python, ZeroMQ – for master-slave node communication, GlusterFS as common filestore, MongoDB for statistics, NodeJS – for web based dashboard
  - You can start with one conversion machine and scale to hundreds

Future work

- Support self-archiving (two phase cataloguing)
- Include CRIS functionality – through a separate CERIF based system
- Various new features for presentation of content

Thank you for your attention!

- More info:

iostath@ekt.gr

nhoussos AT ekt.gr

pstath@ekt.gr