Building an open source based, open standards, infrastructure for the large scale provisioning of reusable open content

Π. Σταθόπουλος, Ν. Χούσσος, Γ. Σταύρου, Ε. Σαχίνη, Ι.Ο. Σταθοπούλου, Κ. Σταμάτης, Α. Σουμπλής

email: pstath;gstavrou@ekt.gr

Εθνικό Κέντρο Τεκμηρίωσης/Ε.Ι.Ε.
National Documentation Center /N.H.R.F.

EKT: The National Documentation Centre

“Ανοιχτά Πρότυπα & Ανοιχτό Περιεχόμενο. Τα θεμέλια της ανάπτυξης του Internet”

“Building an open source based, open standards, infrastructure for the large scale provisioning of reusable open content”

- “**reusable**” and “**open**” content:
  - functionality ?
  - standards ?
  - software ?
  - emerging requirements ?
- An open source software **stack** for delivery of
  - reusable, open content
  - in large scale, using open standards
  - and addressing emerging requirements

Content types and characteristics (I)

- Different content types but similar functional requirements, examples:
  - **Research output**, e.g. papers, reports, patents, documentation, conferences, books, data: **didaktorika.gr**, **helios-eie.ekt.gr**, etc.
  - **Cultural - Humanities repositories**: from research papers to commented cultural artificats: **pandektis.ekt.gr**
  - **Educational – reference material**: studies, books, multimedia courses, project outcomes, e.g. **repository.edulll.gr**
  - **Digitised and/or born digital e-books**: **ebooks.serrelib.gr** (*pre-release*)
  - Library automation systems and OPACs: linking the digital with the physical realm, **abekt.gr**
  - ***Digital objects in general***

Content types and characteristics (II)

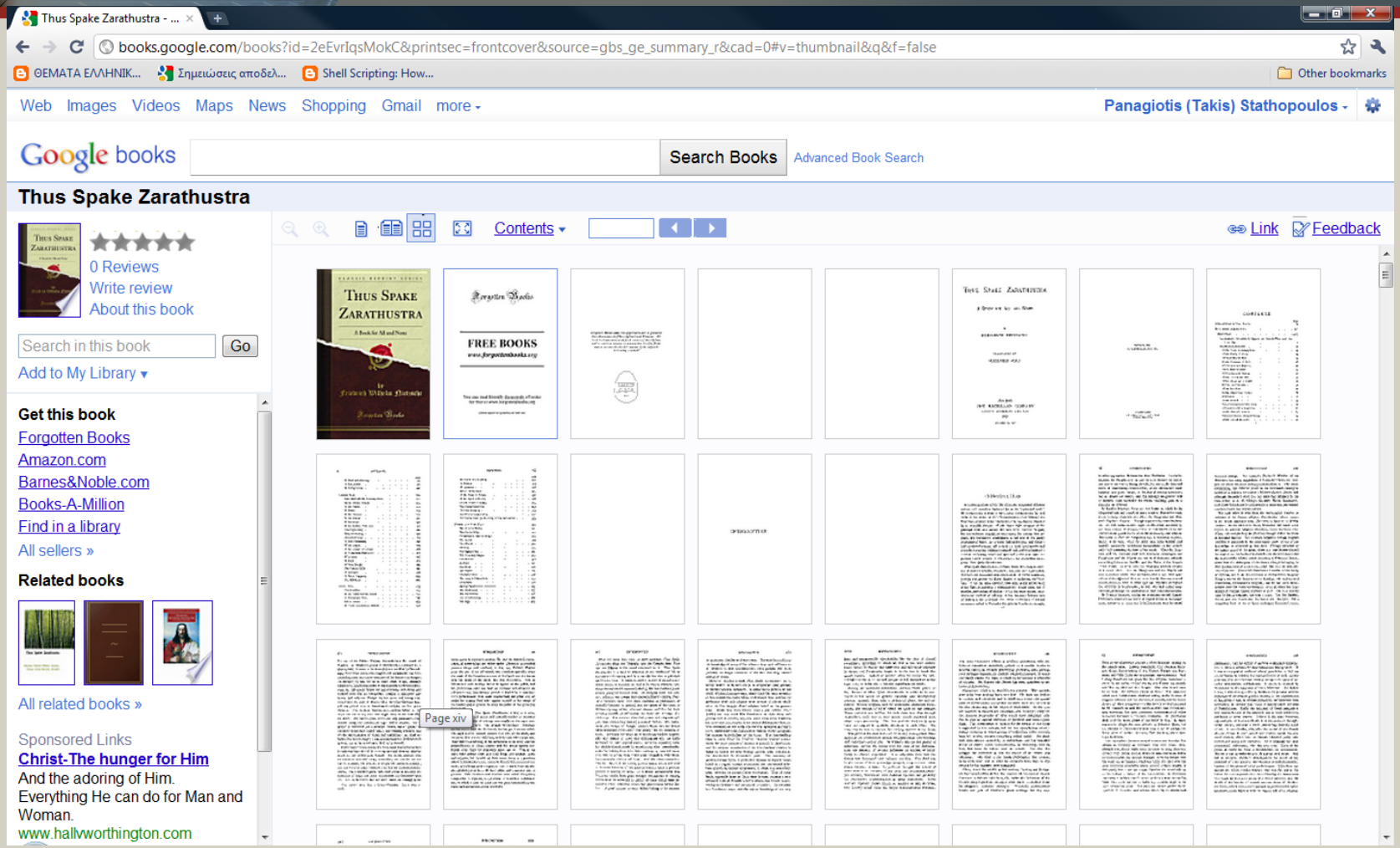
- **Common characteristics:**
  - Online databases providing access to ***digital to objects*** (e.g., books, articles, images, mutlimedia) accompanied with **rich *metadata***
  - Collection, dissemination, preservation of material that has *archival* ***value for future reference***
- ***Different requirements and emphasis in features implemented in comparison to*** Web Content Management Systems
- ***Fast evolving landscape. Full range of features required:***

  copyright information management, ***metadata-bibliographic data management****,* ***metadata organisation and interoperability mechanisms****,highly automated backoffice digitisation processes support,digital content interoperability and reuse frameworks,* ***persisted identifiers****,* ***end user interfaces and intuitive content interaction****,* ***highly efficient storage and computing resources management***

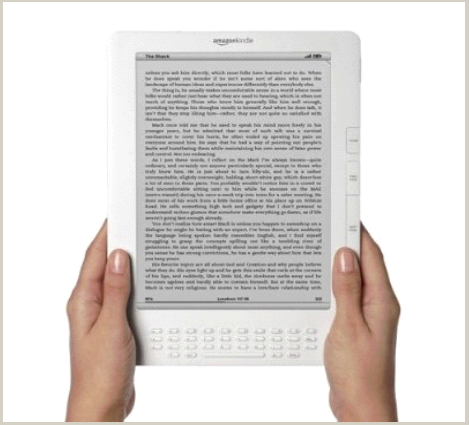
Standards (… so far)

- **Open** and **reusable**:
  - Digital content, file formats, encoders, etc.
  - Metadata description:
    - *UNIMARC*,*MARC21* to *DC* and MODS
  - Interoperability protocols and interfaces:
    - *OAI-PMH*, OAI-ORE, Z39.50, SRU/SRW, *Europeana ESE*, etc.
  - Unique and persistent identifiers
    - Independent from repository software,
    - HANDLE.NET (RFC3652), openURL

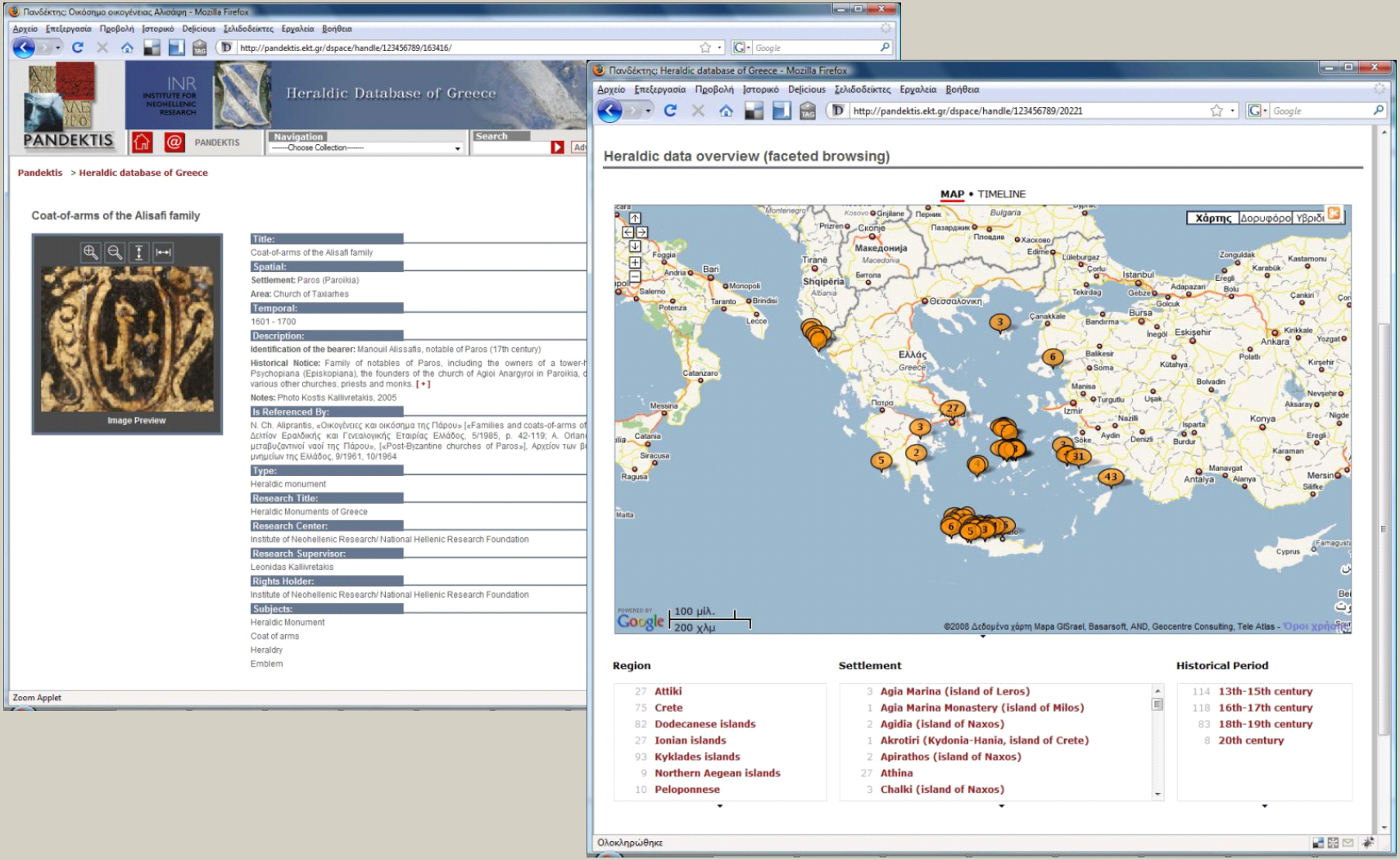
… viewers and enhanced user experiences



… novel reading devices



… mashups and linked data



An open content open source software stack

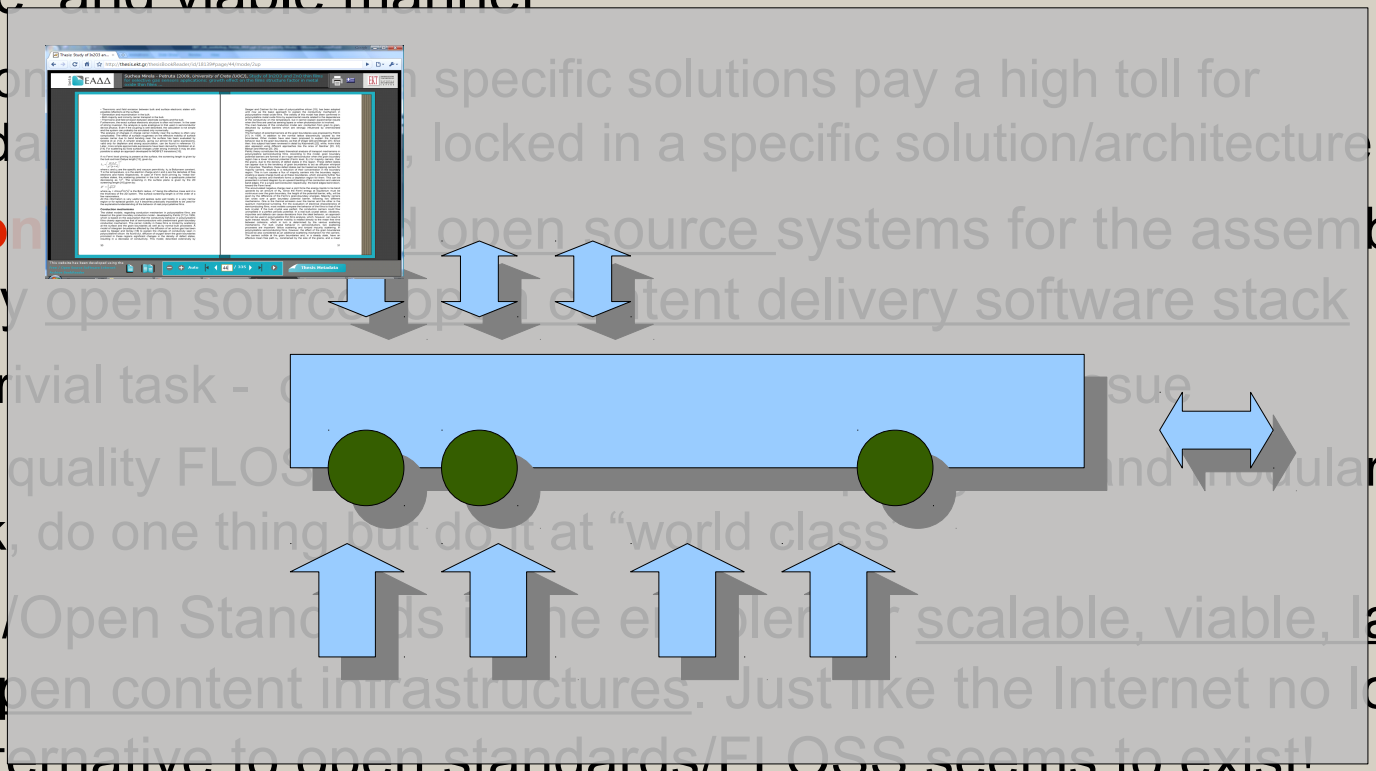
No easy task to tackle the full range of these issues, in a scalable and viable manner

- Custo… specific solutions pay a high toll for stand… locked in a specific solution/architecture

Solutio… …d open source systems for the assem…bly of a fully open source open content delivery software stack

- Not trivial task - …sue
- High quality FLOS… …and modular stack, do one thing but do it at "world class"

FLOSS/Open Stand…ds the enabler of scalable, viable, large scale open content infrastructures. Just like the Internet no long term alternative to open standards/FLOSS seems to exist!



The open software stack components

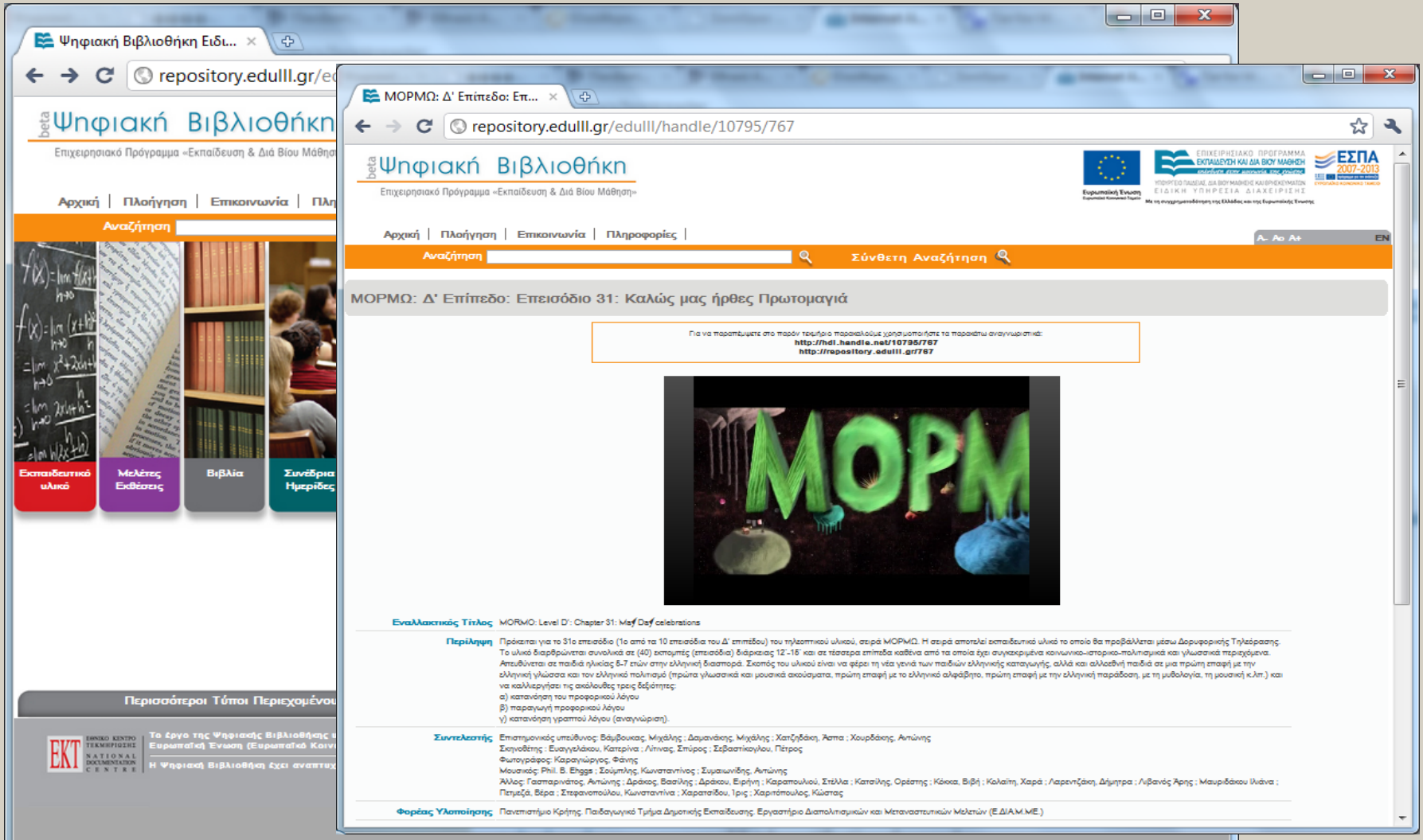
e-OPACs	Digital Repository	High capability viewers	Scalable image servers
Bibl. metadata mgt. (UNIMARC/MARC 21) Interoperability with legacy systems (MARCXML, Z39.50) Connection to the "material" realm (ILS, e.g. ABEKT)	Metadata mgt. (DC, mappings) Interoperability mgt. (OAI-PMH, ESE,...) Persistent identifier (with LHS) Digital object mgt.	Enhanced intuitive user experience Page be page online reading Arbitrary thumbnail views Arbitrary zoom in /zoom out levels Rotation and full text search Open standards compliant, no Flash	Dynamic generation of image quality and size levels Open std: TIFF,JP2000,.. Digital material mgt OpenURL capable Full text search support (OCR XML) Highly scalable archs.
PAPZAR2, ZEBRA	**DSPACE + FEDORA = DURASPACE**	**IA Open Library BookReader**	**IA Open Library Web and Datanode server**

best of breed: open source, setting the standards pace, world class features, scalable and robust communities

Case studies

- Examples of “what can open source do for you”
- 3 case studies:
  - The Education and Life Long Learning Repository
    - Persistent Ids, rich user experience/custom tag bilingual tag clouds based on documentation metadata
  - The e-OPAC “alpha” version of openABEKT
    - Connecting the digital with the physical realms
  - Public Library e-books repository pilot: the Serres Public Library pilot
    - Backwards interoperability, repositories, enhanced viewers and image servers infrastructure

repository.edulll.gr



ΑΒΕΚΤ e-OPAC

Οτιδήποτε

Αναζήτηση

Επιλεγμένα Βιβλία

Θέματα

Αρχαία Ιστορία (11)
Ελλάδα (33)
Ελληνική ποίηση (75)
Ημερολόγια (32)
Ιστορία (43)
Μακεδονία (58)
Νεοελληνική Λογοτεχνία (12)
Ξενόγλωσσα (55)
Παλαιό (23)
Παλιό (69)

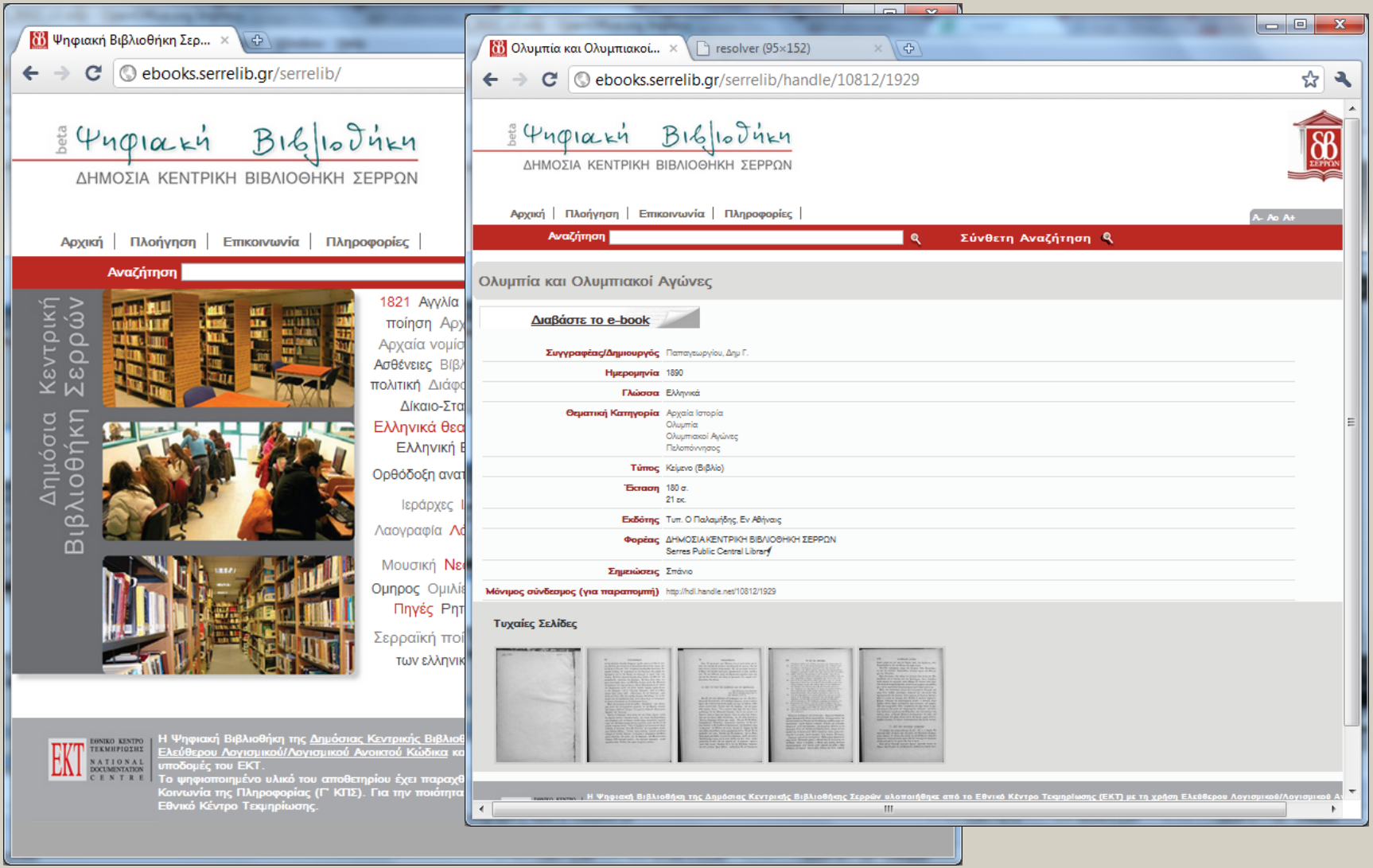
Περισσότερα...

Συγγραφείς

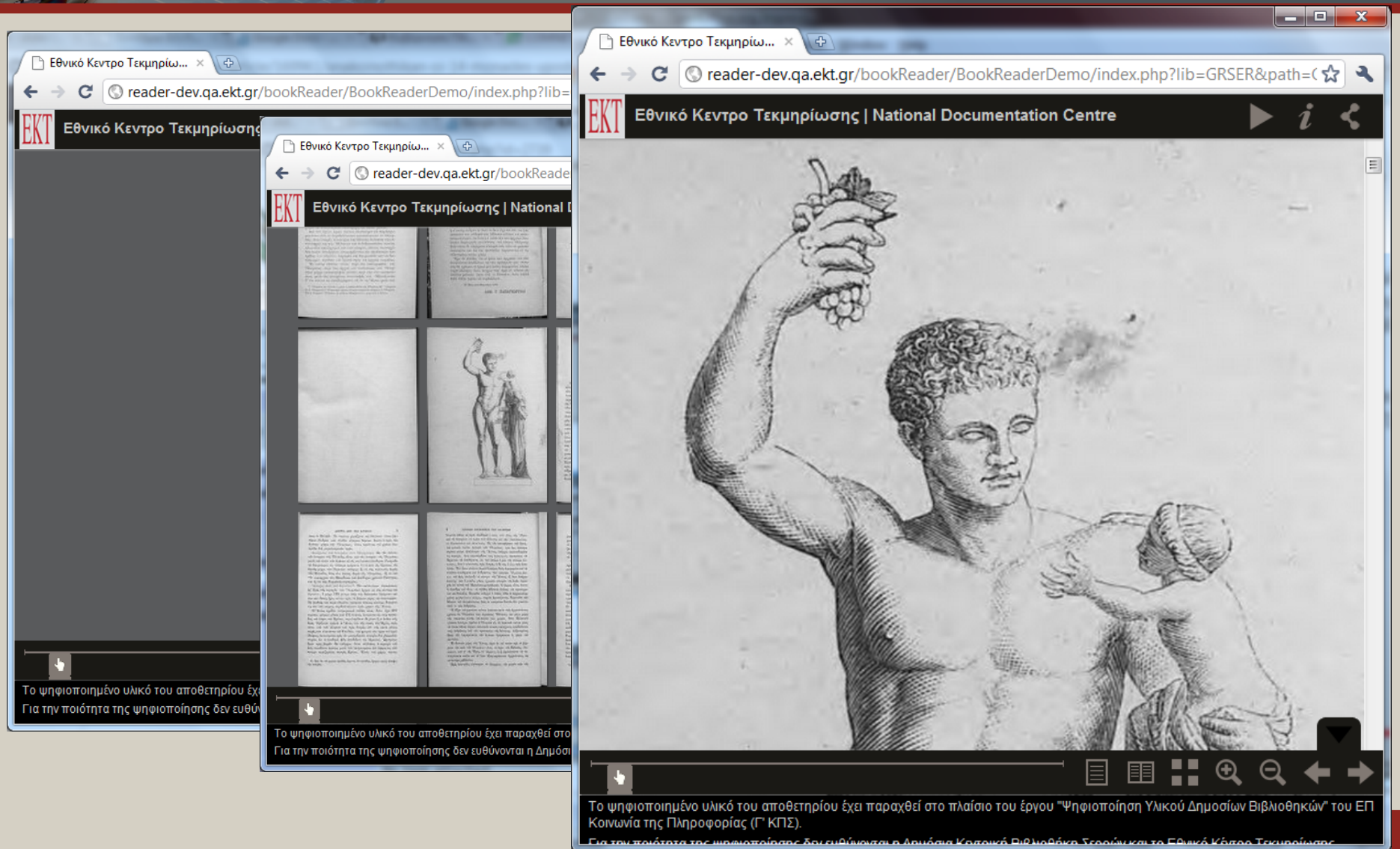
Darwin Charles (2)
Pellico Silvio (2)
Αγγειοπλάστης Γεώργιος Κ. (9)
Αργυρός Αθανάσιος (2)
Γρηγορόπουλος Μιχαήλ Σ. (2)
Δαμαλάς Νικόλαος Μ. (4)
Ευριπίδης (3)
Ζωγράφος Ιωάννης Σ. (3)
Καζάζης Νεοκλής (2)
Καφταντζής Γιώργος (16)

Περισσότερα...

http://ebooks.serrelib.gr



http://hdl.handle.net/10812/1929



Enhanced image viewer/server infrastructure

Internet Requests

Frontend
Load Balancer

Backend Processing

Shared Storage

EKT's virtualised infrastructure

webnode(s)
pool

datanode(s)
pool

Data

Shared Storage: 1 Master
(RW) + N “slaves” (RO)

Currently: 300K pages, 3M imminent from
didaktorika.gr, 10 resolutions/page

Characteristics

- Caching
- Load Balancing
- Error compensation through VCL
- Backend Health Check

- Lightweight & Simple
- Stable
- Right for the job

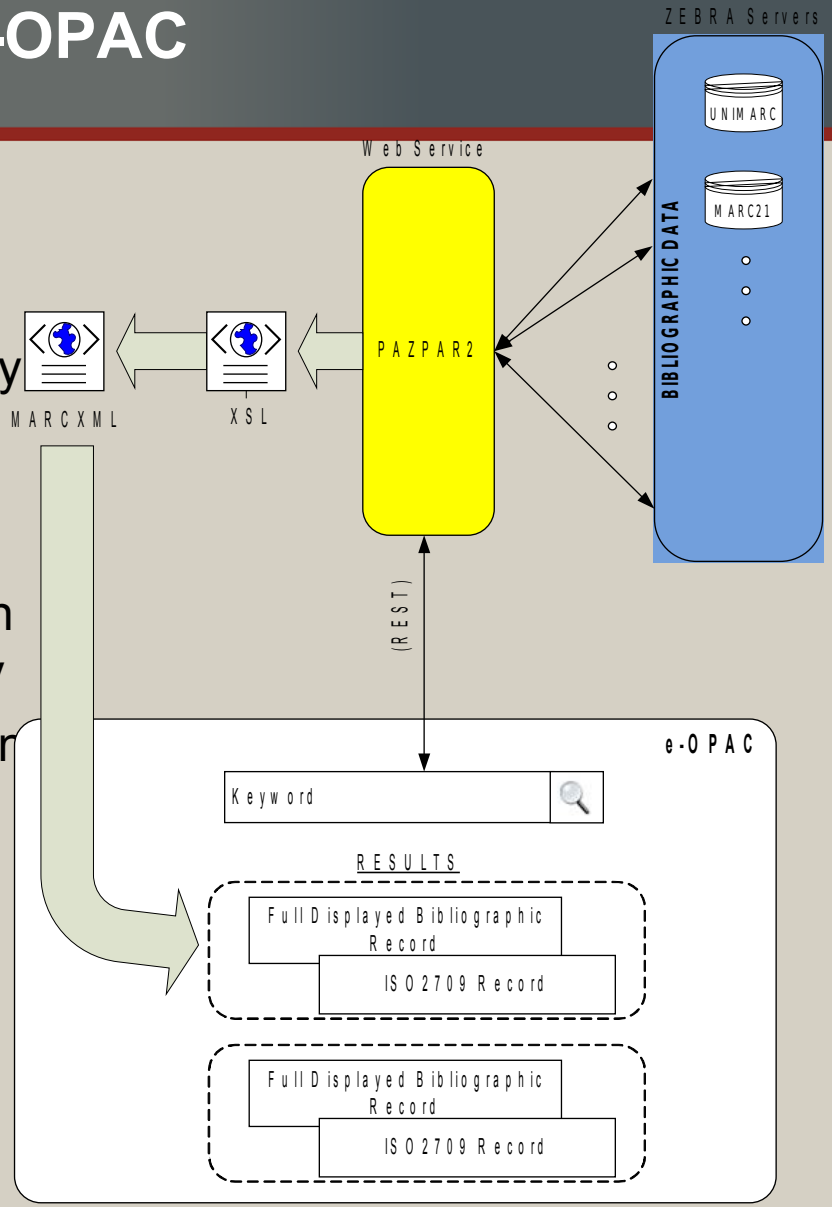
- On the fly Image Processing
- No need for several Images with different resolutions
- Open and Extensible API

"Infinite" Scalability
Stateless
Highly Available
End to End FLOSS
Layer Security

Apache Tomcat/DSpace, NGINX and Varnish cache engine in high availability configuration, PostgreSQL DB, on CentOS VMs

(open)ABEKT e-OPAC

- ABEKT new e-OPAC
  - Different need: manage the library (digital and physical) using bibliographic tools and a librarian approach
  - The interoperability layer between the physical and the digital library
  - “Richer” metadata and information
- ROR over OSS components



Digital repositories software

- Dspace 2.0: with Fedora inside
  - EKT participates at the design team
- Dspace + Fedora = Duraspace
- Advanced content provisioning and preservation functionalities
- The core element for an interoperable repositories infrastructure
- Preparing for operation as SaaS and over IaaS Cloud
  - Delivers as SaaS a complex service

Future work and informational links

- Current and forthcoming work
  - Viewer with full text search: OCR XML
  - Back port features to existing repositories
  - Integrate to OJS the page by page functionality
  - openABEKT
  - Automate already existing economies of scale: offer SaaS capabilities for the whole range of tools
- Links
  - http://www.duraspace.org/
  - http://openlibrary.org/dev/docs/bookreader
  - http://www.indexdata.com/zebra

Thank you for your attention!