The role of digital long-term preservation in the eSciDoc project

Natasa Bulatovic,
Max-Planck Digital Library
bulatovic@mpdl.mpg.de

eSciDoc: Background information

- Joint Project between Max Planck Society and FIZ Karlsruhe
- Funded by German Federal Ministry for Education and Research (BMBF) until Mid 2009
  - Additional substantial own efforts by both partners
  - Both Partners committed themselves to ongoing efforts until at least 2011

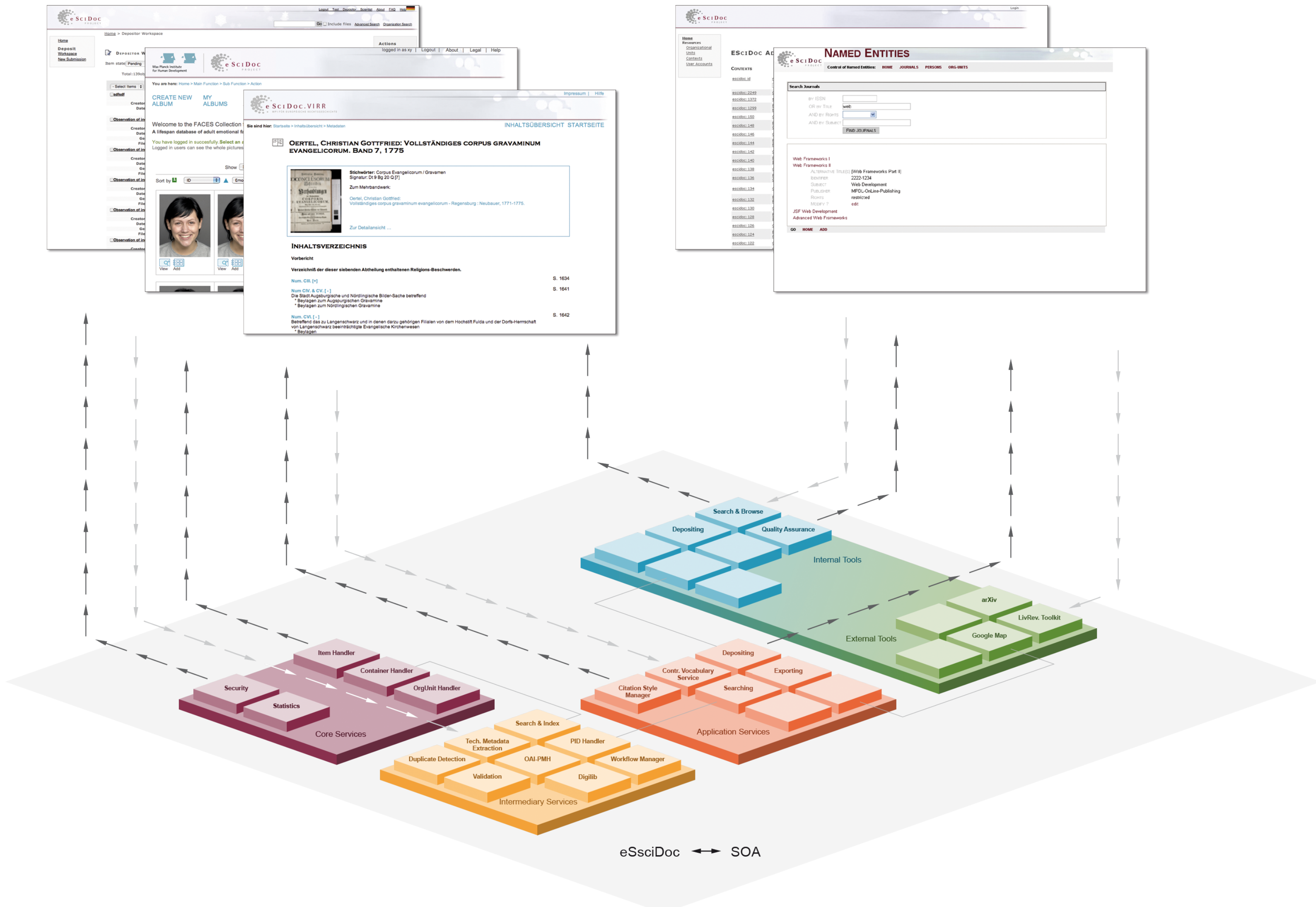
eSciDoc: what it provides?

- **Organizational aspects**
  - ✓ Business processes: introspection into the “scientific information and scientific data life-cycle”
  - ✓ Modelling of the “scientific workflows” (usage, scenarios, use cases, process worfklows)
  - ✓ “SOA” is not only technical infrastructure - it is driven by organization-wide processes and requirements

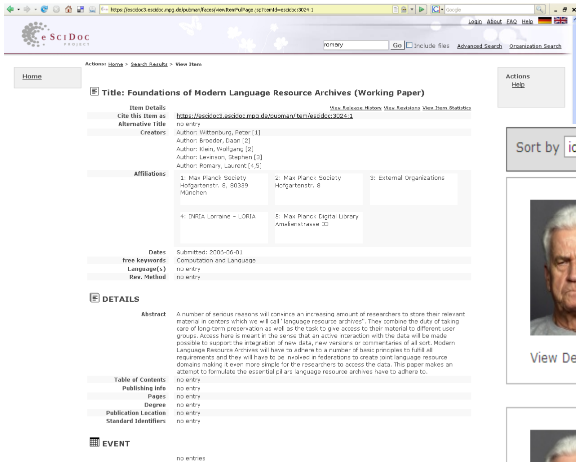
- **Technical aspects**
  - ✓ service infrastructure instead of silos applications
  - ✓ solutions to visualize, publish, relate and manage data
  - ✓ easy composition of services and data mashups and repurposing

Both organizational and technical activities facilitate the growth of the organization and the dissemination, accessibility and easy reuse of the scientific results across disciplines

eSciDoc project landscape



eSciDoc Solutions



Publication management

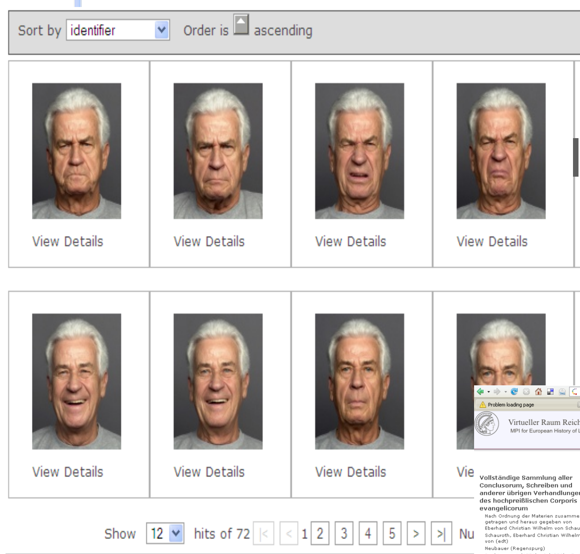
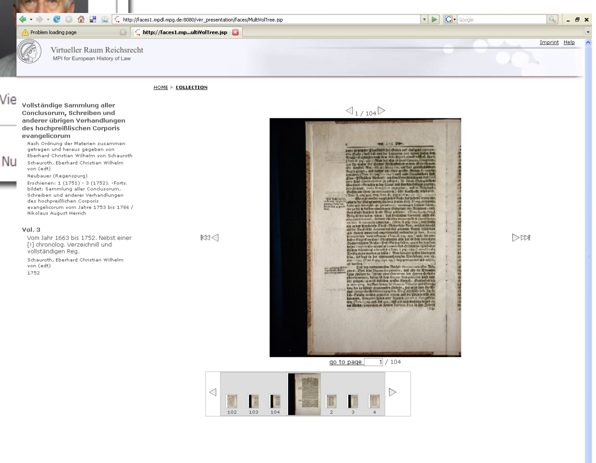


Image collections



Scanned books

Challenges

And facts for eSciDoc project related to the digital preservation and long-term archiving

Challenges and facts: Data structure, semantics and descriptive metadata

- *Challenge: Support for variety of data (publications, old manuscripts, microscopic images, patents, datasets)*

✓ High level of abstraction in data modelling and implemented data services **(items, containers).xml**

✓ Specialization through **content models** gives **contextual** information on what data it is (publication, scanned book page, lexical resource, collection of “digital things”)

✓ OWL ontology modeling (in progress) - interoperable content models

✓ Metadata schema use elements from known namespaces

✓ a service validates data structurally and semantically based on rules defined by the community (Schematron)

Challenges and facts: communities and workflows

- *Support for variety of workflows, enable their configuration by declaration from the users and respective community*

- ✓ Workflow actions (events) are used in various workflows
- ✓ Workflows differ not only by types of resources but also by community
- ✓ Core life-cycle for each resource (pending, submitted, released, withdrawn)
- ✓ Users decide on their specific workflows e.g.
  - publish data imediately with their deposit/ingest (allow quality assurance and metadata enrichment later)
  - publish data after initial quality assurance and metadata enrichment activity (allow quality assurance and metadata enrichment later)

- ✓ Workflows also define user roles and privileges in a specific administrative context

Challenges and facts: object identification

- *Different communities decide what to identify and when in a different manner and at a different state of the completeness of their data*

- ✓ Each resource (metadata, files) **must** be identified latest before its first time "publishing"
- ✓ Users may decide what to identify (resource, resource version) and when to assign the persistent identifier (during creation, updates, publishing)
- ✓ Users may configure when to assign the persistent identifier on a repository level (planned: extend these possibilities depending on content model and applied data management workflows)
- ✓ System understands and keeps all persistent identifiers of the resource (if such existed before the resource has been created in the system)
- ✓ System in addition generates a persistent identifier for each resource and resource components (i.e.files)

Challenges and facts: object identification

- But which PID system to choose?
- The service is designed to support different persistent identifier systems. Different methods to generate identifiers could be selected (serial, random or semantic).
- Simple REST interface
  - GET: /<service>/<prefix>/<suffix> - resolve
  - PUT: /<service>/<prefix>/<suffix> - registration at PID system
  - POST: /<service>/<prefix>/ - identifier generation and registration at PID system
  - DELETE: /<service>/<prefix>/<suffix> - delete

Challenges and facts: metadata and extraction

- Resource description: All metadata are in XML and use known and publicly available namespaces wherever possible
  - Item: escidoc:3914
  - faceItem: escidoc:6415
- Files: descriptive metadata and in addition technical metadata
  - Technical Metadata Extraction service, based on the JHOVE.
  - PRONOM identifier is associated with files.

Challenges and facts: provenance data

- Support for event logs and version history (what happened with the content?)
- PREMIS v1 created for each resource (events info)

Item: escidoc:3914

Challenges and facts: software, standards, documentation

- Software
  - ✓ all used software is an open source software
  - ✓ eSciDoc software available under CDDL license (open source)

- Documentation
  - ✓ Accessible (still a lot of work to do)
  - ✓ Specifications and some design are available via MPDL colaboratory e.g. http://colab.mpdl.mpg.de/mediawiki/PubMan_About

- Standards and formats
  - ✓ XML, XSD used for interface and service operation messages
  - ✓ Use of metadata standards
  - ✓ PREMIS
  - ✓ Open formats wherever possible e.g. metadata, lexical resources TEI-XML etc.
  - ✓ Containment: Fedora XML objects for content resources
  - ✓ OAIS reference model – compatible
  - ✓ OAI-ORE model (for content resources dissemination, but also to describe AIPs)

Bitstream preservation ready ☺

Lessons learned, next steps

- Have we covered all possible aspects? Has anyone? Many efforts!
- long term and digital preservation aspects must be considered from the very start in infrastructural design and implementation
- Things are getting more complicated in dynamic environment (data are changed often)
- Selection process: preserve all? Who decides? How well defined is the border between “validity and needed data” and data we need to preserve for “institutional memory” e.g. “the making of..”?
- As many copies as possible => data interoperability => sharing and replication of data as much as possible – redundancy does not hurt (it costs ☺)
- Collaboration and cooperation with other projects necessary (its not a single unit effort)
  - Building the eSciDoc community – Special Interest Group LTA (SIG)
  - Nestor – working group
  - SUB Göttingen – Kolibri for ingestion and automatic extraction of metadata for LTA

Where to next?

- Most important issues ☺
  - ✓ Organizational commitment and policies
    - ✓ MPS aspect: Centralized and decentralized efforts
  - ✓ Sustainability in the organizational commitments and policies
  - ✓ Long-term-ability of the organizational commitments and policies

Thank you!

- bulatovic@mpdl.mpg.de

Resources

- eSciDoc project web pages, Infrastructure download
- http://www.escidoc-project.de
- MPDL collaboratory network
- http://colab.mpdl.mpg.de
- MPDL software download
  - http://escidoc1.escidoc.mpg.de/projects/common_services
  - http://escidoc1.escidoc.mpg.de/projects/pubman/
  - http://escidoc1.escidoc.mpg.de/projects/faces/
  - http://escidoc1.escidoc.mpg.de/projects/virr/
  - http://colab.mpdl.mpg.de/mediawiki/ESciDoc_Admin