Gabriella Ivacs, OSA, ivacsg@ceu.hu
Sofia, September 16, 2009
The PARALLEL ARCHIVE project:
www.parallelarchive.org

• Developed by the Open Society Archives@CEU, funded by the Institute of Record, maintained, powered by an external company and CEU

• OSA Laboratory Project, it is not the OSA Digital Repository

• 2007: consultation with archival institutions, scholars, fundraising experts, developers

• 2008: specifications, project planning, implementation, user testing

• 2009: sustainability planning, 3-year roadmap, fundraising, prototype roll-out at small research communities

• 2010: dissemination plan, sustainability plan, governance model and development (batch ingest, more research tools)
PA was created to address the problems scholars face when working with archival material.

**Obstacles scholars face during research**

- **Locating Archival Collections**: It is often hard for scholars to find collections of interest for their research. PA provides information on the physical source of material, helping scholars locate relevant archival collections.
- **Rigid Organization of Archival Collections**: The hierarchical organization of traditional archives makes research difficult for the non-initiated. With its library-type search, PA gives access to individual archival documents.
- **Absence of Archival Content Online**: Archives have been slow to digitize their content. PA proposes to capture scholars' individual efforts to give general access to archival material.
- **Lack of Tools to Study Documents Directly Online**: Scholars' digitized archival material often ends up disorganized on their desktops. PA creates PDFs, offers tools to organize and view them, provides OCR'd text, and allows online page annotations.
WHY DO WE NEED A PARALLEL ARCHIVE?

PA was also created to address the problem of „small” collections, archives, research institutes.

Obstacles Small „archives” face

- LACK OF INFRASTRUCTURE AND TOOLS
  Small institutions, research centers, emerging archives, such as documentation centers are often under-funded. They cannot invest into infrastructure, no strategic approach.

- ARCHIVAL WORK IS NOT A PRIORITY
  They are more concerned with publishing research results, delivering records to their clients. Unaware of basic archival principles: provenance, chain of evidence, authenticity.

- SPECIAL TYPE OF CONTENT
  Multilingual collections, mixed format, obsolete media, unclear legal status offers a complexity of issues hard to tackle.
Parallel Archive is a space where researchers can upload, study, and share their scanned archival sources. It facilitates scholarly research, collaboration, and a critical approach to documents.

It is at once

- **A DIGITAL ARCHIVE**
  - PA is a digital repository made by end users where you can upload and search for documents

- **A PERSONAL WORKSPACE**
  - PA is a personal workspace where you can store your documents, work with them directly online, and access them anywhere through the internet

- **A COLLABORATIVE NETWORK**
  - PA is a collaborative environment for scholars where you can discuss documents and general research themes
WHAT SORT OF DOCUMENTS CAN YOU FIND IN THE PA REPOSITORY? (Collection Policy)

- documents of historical value, testimonies of the past
- original documents, somehow unique, generally non-published

EXAMPLES:
- official, legal, and organizational records and reports
- manuscripts, letters, journals
- photos, maps, prints, plans
- newspaper clippings, annotated publications
- ephemera and grey literature

Search for individual documents by type, subject, name, language, country, date, source, etc.

Locate collections of interest relevant to research through information on original sources.

Collect and store documents of interest at PA personal workspace.

HOW TO INCORPORATE ARCHIVAL SOURCES INTO RESEARCH?
2: USEFUL WORKING TOOLS FOR RESEARCH

Tools to work with archival materials

- **ORGANIZE DOCUMENTS**
  - 1. Hierarchical system: folders and sub-folders.
  - 2. Tagging as alternative access

- **EDIT TEXT**
  - To improve, to share, to work with text version of documents.

- **ANNOTATE PAGES**
  - To annotate documents page by page as you read them.

- **EMAIL AND DOWNLOAD**
  - To e-mail a citation or to download any document to computer.
CREATE A DOCUMENT ON PA (1)
Upload your document

Choose your upload method

Browse and... upload your archival material
CREATE A DOCUMENT ON PA (2)

Describe your document

Give as much information on the document as you can and be sure to fill in all required fields!

At the end of this process, you must:
- decide whether to keep your document private or make it public
- submit your document
3 : PERMANENT URLs FOR ACCESS AND CITATION

PERMANENT ACCESS TO DOCUMENTS
Upload, store, study, and manage documents online anywhere through the internet

A TOOL FOR REFERENCE AND CITATION
Provide direct links to primary sources in publications or papers

CURATED SCHOLARLY CONTENT
PA exercises curatorial control of content to maintain a scholarly standard
4: A NETWORK OF SCHOLARS AND IDEAS

WORKING WITH OTHER SCHOLARS
- Find scholars with similar research interests and collaborate
- Form research groups around specific documents or broader topics

ENRICHING PA SOURCES THROUGH NETWORKING
- Tag documents of interest to enhance access
- Comment on others' documents to share advice or expertise
CONCLUSION: AT PRESENT A TOOL FOR INDIVIDUALS

- Upload and Store Documents
- Organize Documents
- Share Documents
- Search for Documents
- Collaborate with Others
- Edit OCR'd Texts
- Annotate Pages
- Refer to Documents
1.1 Compliance with the OAIS Reference Model
2. Sustainability of digital formats and content in PA
Context: preservation and curation

- **Digital formats, files:**
  - Original files: TIFF, JPEG, PDF
  - Surrogates: JPEG, PDF, HTML
  - Metadata: Dublin Core and Technical metadata saved

- **Authenticity, provenance:**
  - Hash code is added to the original file
  - Source is linked back to archives, contextual information is to be found

- **Risks:**
  - Quality of master file is not good
  - Preservation metadata schema needs to be developed
  - Versioning is needed
  - Master files need to be exported to other repositories

- The provenance is defined by the user, forgery?
- Archival finding aids are not item-level, digital object cannot be identified
3. Sustainability of digital formats and content in PA

Context: preservation and curation

**Interoperability of hardware, software:**

- PHP programming language, disclosed source code
- Open Formats: TIFF, JPEG, PDF
- OS solutions except for OCR
- PostgreSQL Database
- Linux Red Hat configuration
- Technical Documentation is almost ready

**Risks:**

- Eliminate dependency: OCR
- Instead of Linux Red Hat, Debian GNU Linux?
- Complete documentation!
4. Sustainability of digital formats and content in PA
Context: preservation and curation

- **Access:**
  - Unique persistent identifier, Handle.net System
  - Private and public digital objects and metadata levels

- **Digital Rights Management:**
  - Copyright Policy, Privacy Policy and Service, Code of Conduct
  - Strong Moderation, procedures
  - Collaboration with archives
  - Planned: Legal Curation Center, Anonymisation Tool

- **Risks:**
  - Privacy violation
  - Copyright violation
  - Violating restrictions by donors
5. Sustainability of digital formats and content in PA
Context: preservation and curation

- **Organizational Viability:**
  - OSA negotiates with a consortium of research institutes, archives to store master files in their repository
  - Assigned PA team within OSA with clear roles
  - Solid organizational background: OSA is part of the Central European University.
  - Business Plan, Dissemination Plan, Fundraising Strategy is in progress

- **Risks:**
  - Current data and content is not packaged yet to be preserved
  - Small organization: lack of expertise, dependency on external developers
  - Financial stability of the project
6. Preservation and Curation Challenges in the case of Parallel Archive

- User driven system, not institution focused: “folksonomy”
- Dynamic content
- Dynamic system
- Authenticity
- The trusted repository model is hard to apply
- Data Management Plan Content Checklist: too theoretical
- “Mixed” digital content: quality of image, themes, metadata,
- Metadata enrichment: tagging, annotation, OCR, comments, forum posts
- Multilingual content
- What to preserve?