Better Humanities Research in the Network

Providing Context

Tobias Blanke (King’s College)
Overview

- Background of the technical work in DARIAH
- Overview of current situation
- Work on the infrastructure
Background
The mission of DARIAH is to enhance and support digitally enabled research across the humanities and arts.
Data-centric Collaboration

- Data Centric Research: large, rich, and complex
- Design interaction around data
- Scholarly lifecycle perspective

Dave de Roure: New e-Science Keynote

www.dariah.eu
Issues with Integrating Humanities Data

- Qualitative human based data needs novel methods of selection (unstructured)
  - Databases rarely follow standard database schemas
  - Data not just fuzzy but contradictory
  - Data not only incomplete but incompletable
- Few standard formats or interfaces
- **Problems of Curated Databases?**
  - Semantics barrier: complexity and context dependency of research material
    - Many structural and semantic relationships both internal and contextual
  - Resources are strongly based in communities

www.dariah.eu
Connecting Communities

Provide a trusted intermediary that makes content both durable and usable with a “Chinese menu" of added-value services.

DURAPSPACE

Embrace the web (L. Carr)
Currently: No web

Embracing the web
<table>
<thead>
<tr>
<th>DARIAH Partner</th>
<th>New to Digital Arts &amp; Humanities</th>
<th>Project Centre for Digital Humanities</th>
<th>Project Hub for Digital Humanities</th>
<th>Towards Research Infrastructure Hub</th>
<th>Infra-structure and Service Hub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academy of Athens (AA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Archaeology Data Service (ADS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ADONIS (CNRS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Centre for e-Research (CeRch)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data Archiving &amp; Networked Services (DANS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Digital Curation Unit (DCU)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>European University Cyprus (EUC)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fondazione Rinascimento Digitale (FRD)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Götttingen State &amp; University Library (UGOE)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institute for Contemporary History (ICHI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Irish Research Council for the Humanities &amp; Social Sciences (IRCHSS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LaPa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max Planck Gesellschaft (MPG)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nordisk Forskningsinstitut (NFI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rudjer Boskovic Institute (RBI)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>INSTITUTION</td>
<td>SERVERS</td>
<td>DIGITAL MATERIAL</td>
<td>DATABASE</td>
<td>REMARKS</td>
<td>STANDARDS</td>
</tr>
<tr>
<td>-------------</td>
<td>---------</td>
<td>-----------------</td>
<td>----------</td>
<td>---------</td>
<td>-----------</td>
</tr>
<tr>
<td>ILNEo</td>
<td>RAID: 750GB+2 TERRA &amp;</td>
<td>20.000---&gt; 3.000.000.000</td>
<td>ACCESS: EXCEL, XML</td>
<td>LINGUISTIC ARCHIVES</td>
<td>NONE---&gt;DUBLIN CORE</td>
</tr>
<tr>
<td>HFRCo</td>
<td>WIN2003-SRV-SRV-3 TERRA &amp; 1 TERRA</td>
<td>20.000--DOCS &amp; IMAGES</td>
<td>MS-SQL2005</td>
<td>scientific-papers, abstracts, posters</td>
<td>CRMCDOC, XML</td>
</tr>
<tr>
<td>KEINE&amp;KEMNEo</td>
<td>WIN2003-SRV-SRV-SRV-SRV-SRV-SRV-SRV-1 TERRA &amp;</td>
<td>35.000--IMAGES (20.000-RECS)+10.000-RECS</td>
<td>MS-SQL2005</td>
<td>HISTORICAL ARCS COLLECTIONS++ PROSPHOGRAFY DATABASE</td>
<td>UNIMARC: XML, ISO2709</td>
</tr>
<tr>
<td>LIBRARYCo</td>
<td>UNIX &amp;</td>
<td>140.000-RECS</td>
<td>GEAC ADVANCE</td>
<td>BOOKS</td>
<td>UNIMARC: ISO2709</td>
</tr>
</tbody>
</table>

- Individual collections but connected neither in a local infrastructure nor to a national one
ADONIS: A national infrastructure

iRods: validation of the sip

Fedora commons
Fedora object

HTTP
HTTPS
API
java
URL

LDAP authentication

iRods: report describing the fedora object

iRods: "put aip"

PAC
Long time preservation

CINES

Pierre-Yves JALLUD

www.dariah.eu
<table>
<thead>
<tr>
<th>DB/Ontologies/Semi-Structured/Unstructured</th>
<th>Storage/Access/Data Quality</th>
<th>Architecture</th>
<th>Warehouse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central Archive: One big Table</td>
<td>Optimized, fast queries</td>
<td>Federation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cleaned Data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complex Schema</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Static snapshots</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extra Copy of Data</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Federation</td>
<td>Current Data</td>
<td>Mediated Federation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Flexible Architecture</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No copying of data</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slower Queries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complex Schema</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No/Little Cleaning</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mediated Federation</td>
<td>Current Data</td>
<td>Federation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Flexible Architecture</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>No/Little Cleaning</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slower Queries</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Little or No Data Cleansing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Potentially many mappings</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Maybe more P2P? With some distributed mappings?
Building the infrastructure
DARIAH-tech services

- **A+H infrastructure services**
  - Generic Infrastructure Services: AAI and PID
  - Collaboration Tools and Connectors

- **Reference software packages**
  - To create a repository that is part of the DARIAH network
  - Resource Creation Tools
  - Simple VRE
  - Utilities for Humanities Computing, e.g. crowd sourcing services

- **Support data quality**
  - integrated helpdesk
  - Data Seal of Approval

- **Federation and interoperability services**
  - Registries: Services, Data, Metadata

- **Support development community**

www.dariah.eu
Demonstrators and Experiments

- Demonstrators (community build their own infrastructures):
  - ARENA: Web-enable a legacy Z39.50 distributed search for archaeological metadata
  - TEI demonstrator: Build a repository and publication platform for TEI resources

- Experiments (the intermediary)
  - Technical Interoperability: Build a virtual repository layer to exchange research objects based on an event based notification model using OAI-ORE and ATOM
  - PID: Cite my data
  - Rights Management
  - Registry/Linked Data: Connect primary resources in humanities
  - Semantic Service Registry: Service registry to serve both technical and research stakeholders
  - EGI: Use the gCube environment to build a Humanities research environment
Context as seen from the network

I have seen this wonderful artefact in the collections at the excavation in England. I wonder whether there is a similar one in Romania.
Arena Demonstrator

Archaeological Records of Europe - Networked Access

Welcome to the ARENA search portal.

This is the search interface of the ARENA project. It allows you to search by selecting a theme and place. You can build your search by selecting a theme and place. Searches are limited to 5,000 hits. If your search exceeds the limit, you will be asked to refine your search.

Available Services:

1. What
   - Choose a type of monument/site from a list of archaeological themes.
2. When
   - You can choose to search for sites within a certain period range.
3. Where
   - Using a map of Europe choose the area you would like to search.
Search

Keyword:
TEI demonstrator

- Exemplary digital archive for e-Humanities
- Validation against TEI-ALL
- Automatic metadata extraction from TEI header
- Default presentation of full text
  - TEI stylesheets, maybe with profiles
- Creation and management of collections
- Publication of collections (source and/or presentational view)
- Unique identifiers (normalised URIs)

www.dariah.eu
Core guidelines, examples, encoding support

Virtual curation helpdesk

Other sources

Create collections, search, download

Primary source Image

Transcription TEI

Upload documents, create collections, search and other sources, make collections public, select presentation

eSciDoc

Validate, transform, disseminate

eSciDoc Infra

Intermediary Services

Basic Services

Core Infrastructure

www.darian.eu
The Grid – Beyond the web?

Multi-disciplinary and sharing
DILIGENT and D4Science: From Virtual Digital Libraries to VRE’s

Community cooperation

VREs: based on shared local computation, storage and generic service and from EGEE

www.dariah.eu
It is not only the data but it is the tools that make it

- It is difficult to generate meaningful links between research data records
- To find the deeper semantics we need human agents
- The aim of the demonstrator is to show how researcher can do this using a standard EGEE based environment

www.dariah.eu
THANKS

tobias.blanke@kcl.ac.uk
Research Infrastructure (RI)

Definition

- Research Infrastructure is defined as “the physical, informational and human resources essential for researchers to conduct high-quality research”

- It includes:
  - Platforms: Tools, instrumentation, computer platforms and facilities;
  - Resources: Software and information resources, including enabling computer systems and communication networks;
  - Human factors: Technical support (human or automated) and services needed to operate infrastructure and keep it working effectively;

Social Sciences and Humanities Research Council of Canada
http://www.sshrc.ca/web/about/council_reports/2004march_e.asp#3

www.dariah.eu
Digital Humanities

Platforms:
- Highly dispersed in terms of organisation and geographic location
- Often small isolated groups without in-house support

Resources:
- Highly heterogeneous data (digital and analogue) with complex interoperability and semantics challenges
- Large variation of resource types and structures
- Challenges mainly data-driven Humanities.

Human factors
- The 10% rule: Only small, highly competitive funds available
- No Techies! Competition between ‘research’ and ‘infrastructure’: Awareness of new digital methods is limited
- Lone Scholar: Resource sharing not always wanted and possible

www.dariah.eu
Research Data / Humanities

- Size matters
  - TB: digitized collections (25 - 100 MB per image, uncompressed)
  - PB: digital movies
  - TB: spoken language (Shoa Archives)
- Complexity matters more

www.dariah.eu