Integrating Structured Data Resources – The LaQuAT project

Tobias Blanke, KCL
Gabriel Bodard, KCL
Mark Hedges, KCL
Shrija Rajbhandari, KCL
Mario Antonioletti, EPCC
Ally Hume, EPCC
Michael Jackson, EPCC

All those hidden away databases

Integrating Humanities Web Resources
From 2004 DOE Data Workshop report:
“... the data management challenge for systems-oriented research is not simply about data volume. More critical is the fact that the data involved are produced by multiple techniques, at multiple locations, in different formats and then analyzed under differing assumptions and according to different theoretical models.”

Background

- Source data: various data resources produced by researchers in classics (databases, XML, SGML)
- Resources not publicly available, or only available via web site that doesn’t allow you do anything but browse.
- Diverse and non-standard formats/schemas.
- Isolated data sources.
- Would be much more useful to researchers if integrated.
- Aim: integrate resources and allow useful processing to be done.
Participants

- King’s College London
  - Centre for Computing in the Humanities
  - Centre for e-Research
- University of Edinburgh
  - EPCC
- Funded by JISC ENGAGE project

Arts and Humanities

Data
Who were we?

- Arts and Humanities Data Service
- Established 1996, funded until 2008
- Distributed structure: managing executive and specialist subject centres
- Mission: collect, preserve and distribute digital resources funded by AHRC

Who are we?

- Centre for e-Research at King’s College London
- Established 2007
- Incorporates staff and expertise of AHDS and other groups such as AHeSSC (Arts and Humanities e-Science Support Centre)
- Continuity, but some change of focus
AHDS Collections

- Highly diverse in type and size: images, text, music, video, databases, multi-media
- Require specialised knowledge
- Rapid expansion: AHDS data size increased 20-fold between 2005 & 2008
- Increasing number of large objects (e.g. video, archaeology scans)
- Growth of corpora due to major digitisation projects (OCR, OMR, etc.)
- Born-digital recordings of current human developments

Museum of London Archaeological Archive
New Survey of London Life and Labor, 1929-1931
London College of Fashion: The Woolmark Company
Imperial War Museum
Designing Shakespeare
Humanities data

- Qualitative human-centric data that needs novel methods of selection
- Diverse: lack of standard formats and interfaces
- Semantics barrier: complexity and context dependency of research material
- Fuzzy, incomplete (and incompletable), inconsistent, inaccurate

Offloading the complexity: Sharing data via download

- Zip up the dataset and put it on a website.
- Pros:
  - Easy for data provider (us 😊)
- Cons:
  - Possible very large download of only small portion required.
  - User has to install data into a local database to use it.
  - Static snapshot.
Taking on some of the work

- How to deal with the complexity of the data
  - Adding work(-flow) to it
  - Let the computer do some of our work

Linking and Accessing Ancient Texts

An experiment using OGSA-DAI
Motivation

- Grid is about sharing resources.
- OGSA-DAI is concerned with sharing structured data.

OGSA-DAI workflows

- Workflows composed of pipelined activities.
- Activities are installed at the server.
- Data streams between activities.
- Activities for data querying, data transforms, data integration and data delivery.
- Allows some computation to be moved closer to the data.
OGSA-DAI Workflow Example

```
SQLQuery
SELECT * FROM Bands
WHERE name = Bangles;
```

```
ObtainFromHTTP
http://www.someplace.org/stylesheets/webRowSetToHTML.xsl
```

```
TupleToWebRowSetCharArrays
```

```
DeliverToURL
ftp://www.musicplace.org/bands/Bangles.html
```

OGSA-DQP

- Distributed Query Processing
- Allows tables in multiple databases to appear as tables in one database. Can do joins and unions over the tables.
LaQuAT Case Studies

- Case study 1 will integrate the Projet Volterra database of late Roman legal texts at University College London with the Heidelberger Gesamtverzeichnis der griechischen Papyrusurkunden Ägyptens (HGV). OGSA-DAI will provide a consistent schema between the two databases.
- Case study 2 will integrate the Projet Volterra database with the Inscriptions of Aphrodisias dataset, which comprises a corpus of inscriptions in EpiDoc XML format. Again, OGSA-DAI will provide a consistent view on the two data sets.
The data resources

- **Volterra**: Access database with Perl script based publication; mainly text-based searches
- **iAph XML database**: XML data source in EpiDoc; overlap in time with Volterra
- **HGV**: FileMaker Pro; German – use views to translate them

Design Decisions

SQL views

Volterra – Access or MySQL

HGV – MySQL

German-English join table

OGSA-DAI

Views

DQP
Lessons learned

Issues

• Queries that really exploit joined-up structured data
• Drivers
  – Migrate the databases into something that can be used
• Problems with the technology
Architecture

Client

OGSA-DAI

HGV

HGV annotations

Volterra annotations

OGSA-DAI (core + SQL views)

Get database schema requests (OGSA-DAI workflows)

OGSA-DAI (DQP query evaluator)

Create resource request (OGSA-DAI workflow)

OGSA-DAI (core + DQP coordinator)

JDBC/ODBC

Vision: Virtual Data Centre

JDBC/ODBC

OGSA-DAI

RSV - MySQL

German-English join table

Volterra - Access or MySQL

...
Queries

- These are research databases: They contain interpretations and uncertain statements
- How can we join them to reduce uncertainties?
- Join queries or set-based ones?

Inconsistency and Incompleteness in Databases

- Global, virtual database
- Independent databases
- We cannot just repair the underlying databases – no permission
- Can we do Joins (not speaking of performance)?
Semantically correct joins

<table>
<thead>
<tr>
<th>A.Place</th>
<th>B.Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antiochia</td>
<td>150</td>
</tr>
<tr>
<td>Antiochia</td>
<td>100</td>
</tr>
<tr>
<td>Sitia</td>
<td>200</td>
</tr>
</tbody>
</table>

- Repairs?
- Collect consistent statements:
  - Table(Sitia, 200)
  - Table(Antiochia; 100) OR Table(Antiochia; 150)
  - Table(Antiochia; X)

Benefits

- **For KCL:**
  - New research questions through the integrated data resources
  - Integration
  - Technical challenges

- **For EPCC**
  - Further development (new services)
  - Realistic requirements
  - Real life data created by real researchers
Next Steps

- Enhance XML support in OGSA-DAI; extend query language to include XPath (AIST Japan)
- Locate and incorporate more datasets
- Investigate more realistic and complex queries across datasets
- Deal with inconsistencies
- Output of results sets (integration into researchers’ work flow)s/sheet

Outcomes

- DARIAH: Digital Research Infrastructure for Arts and Humanities
- Further ongoing grant applications
- Further development of technologies
- Publications and presentations
Questions

http://laquat.cerch.kcl.ac.uk/