

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



Outline

- Background
- Case Study: Linking and Accessing Ancient Texts
- Demo
- Lessons Learned
- Future Work: DARIAH

JISC



engage
Engaging research with
e-Infrastructure

LAQVAT

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

JISC



engage
Engaging research with
e-Infrastructure

LAQVAT

Arts and Humanities

Data



Who were we?

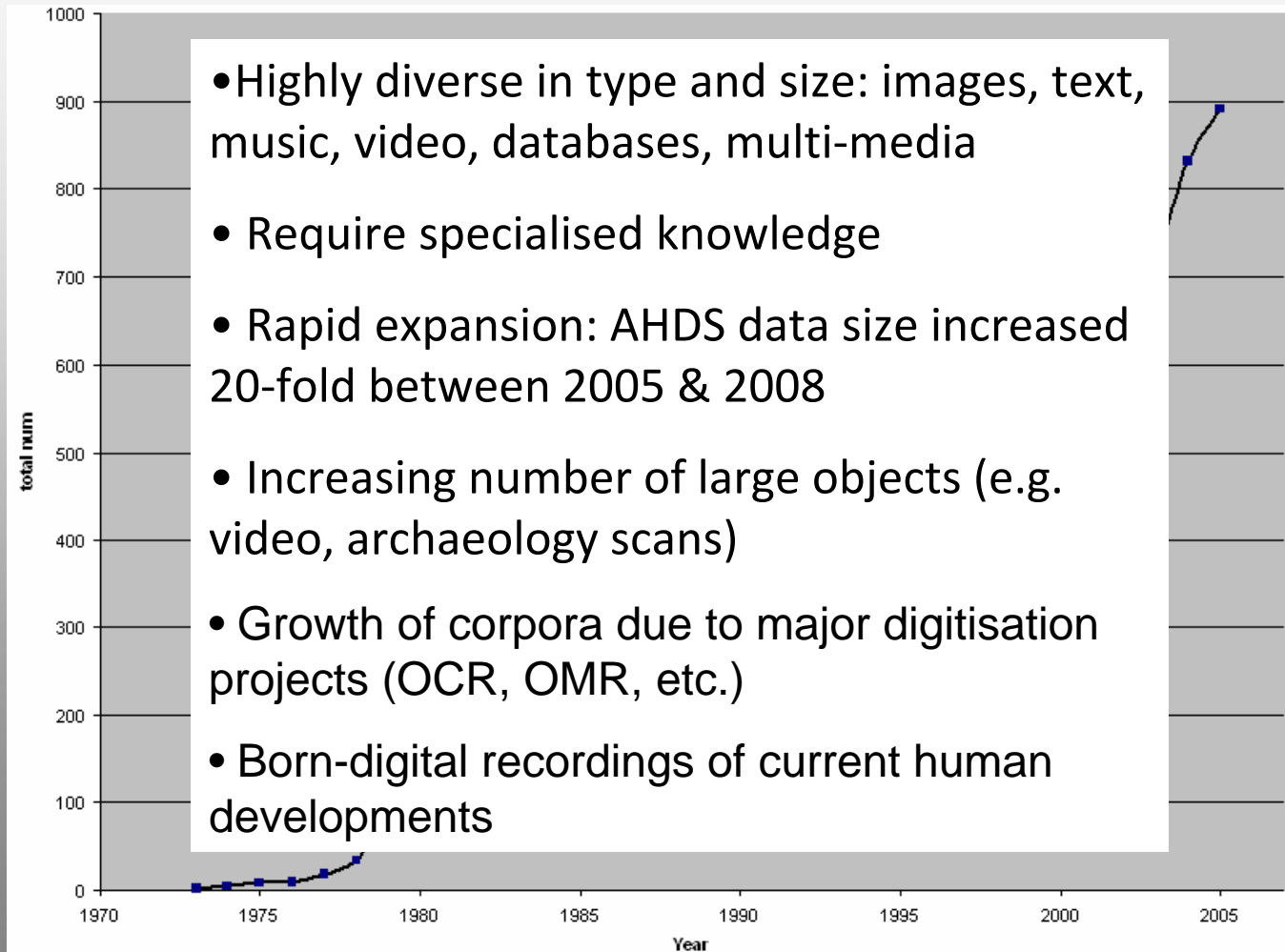


- Arts and Humanities Data Service
- Established 1996, funded until 2008
- Distributed structure: managing executive and specialist subject centres
- Mission: collect, preserve and distribute digital resources funded by AHRC

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





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

JISC



engage
Engaging research with
e-Infrastructure

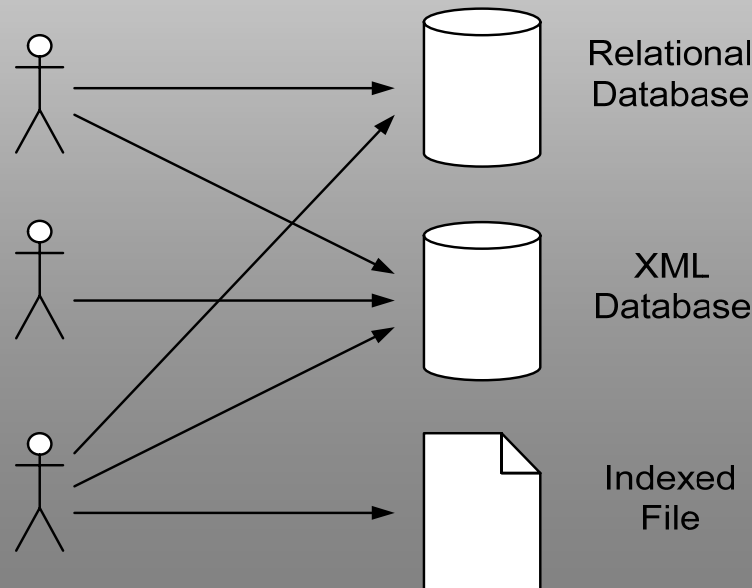
LAQVAT

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

Access

SQLQuery
SELECT * FROM Stuff
WHERE name = stuffy;

ObtainFromHTTP
http://www.someplace.org/styl
esheets/webRowSetToHTML.xsl

tuples

TupleToWebRowSetCharArrays

Transform

WebRowSet XML

XSL

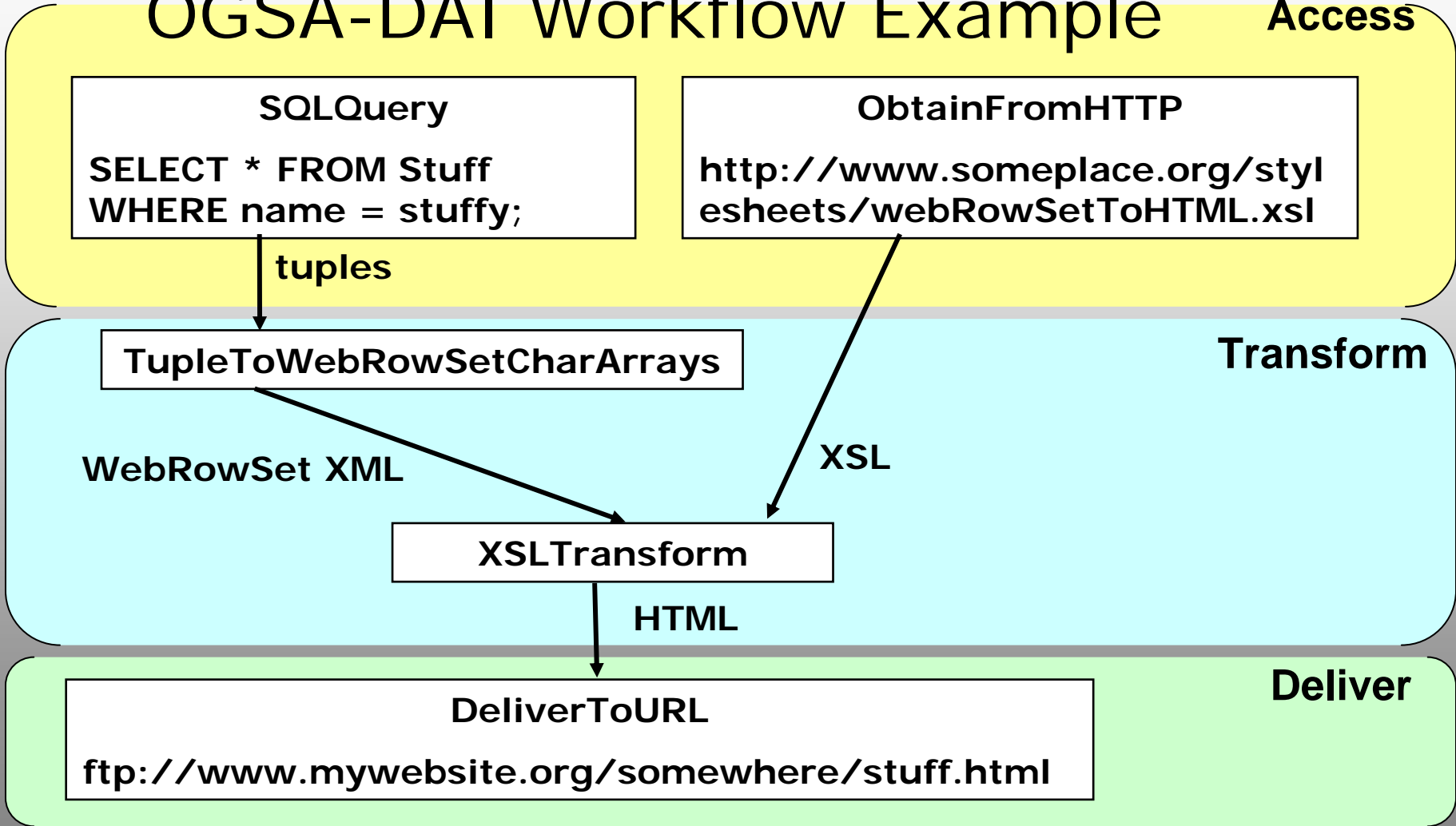
XSLTransform

HTML

DeliverToURL

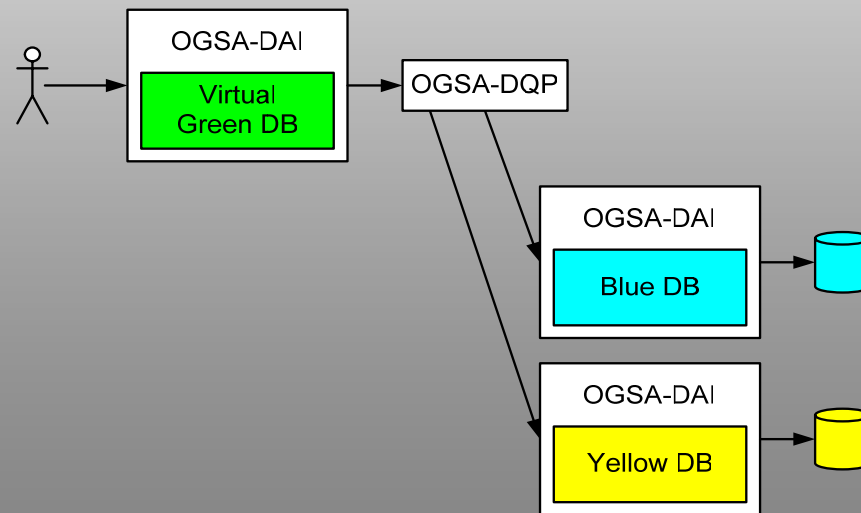
Deliver

ftp://www.mywebsite.org/somewhere/stuff.html

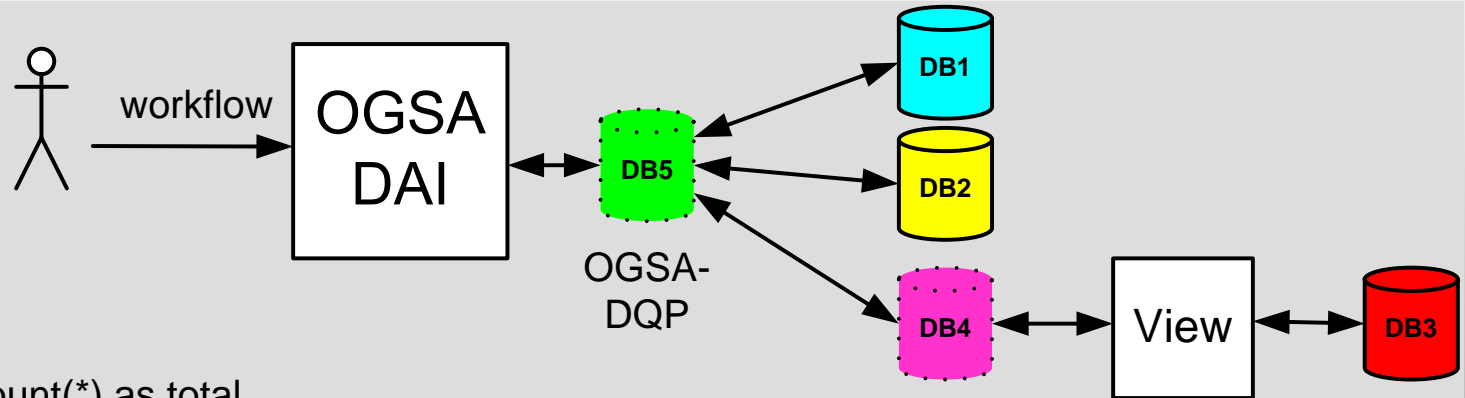


OGSA-DQP

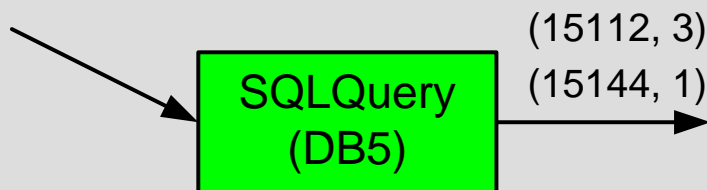
- Distributed Query Processing
- Allows tables in multiple databases to appear as tables in one database. Can do joins and unions over the tables.



CDC Scenario: using OGSA-DQP



```
SELECT zip, count(*) as total
FROM DB1.Cases UNION DB2.Cases UNION DB4.Cases
WHERE Reason = "Flu"
GROUP BY zip
ORDER BY zip
```





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

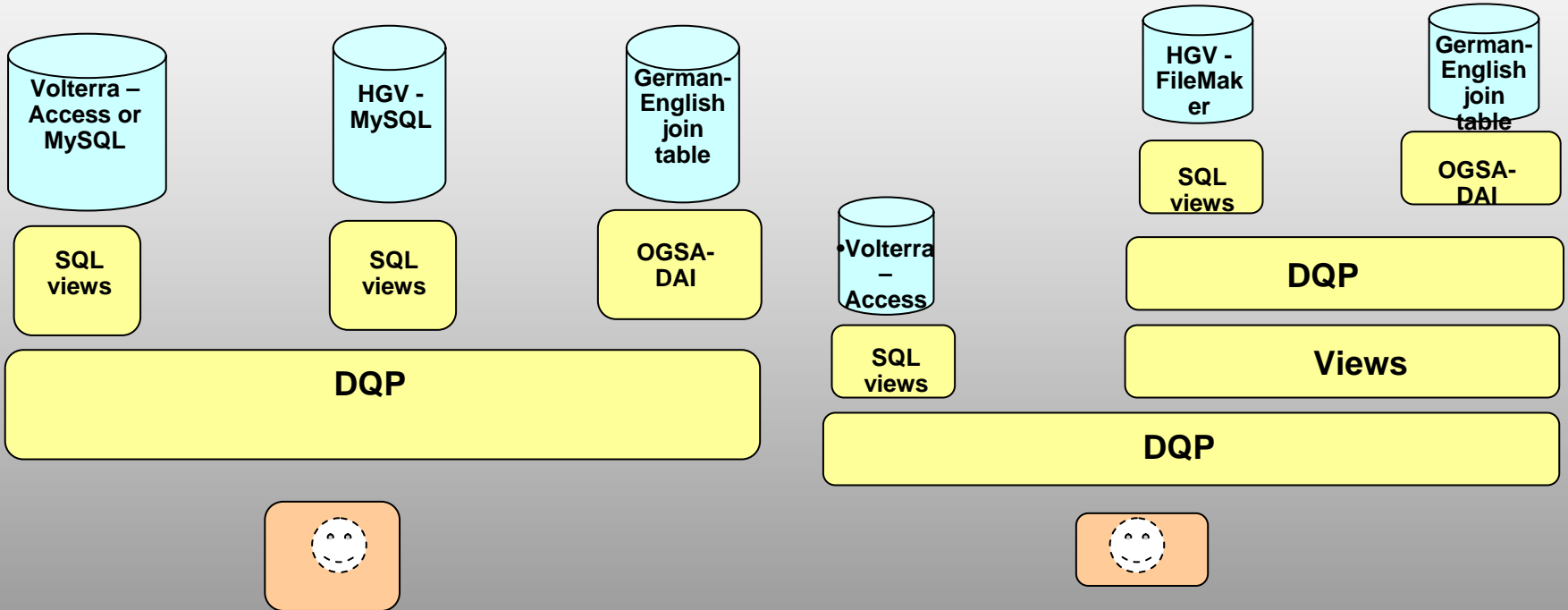


The data resources

- 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



JISC



engage
Engaging research with
e-Infrastructure

LAQVAT

Demo

JISC



engage
Engaging research with
e-Infrastructure

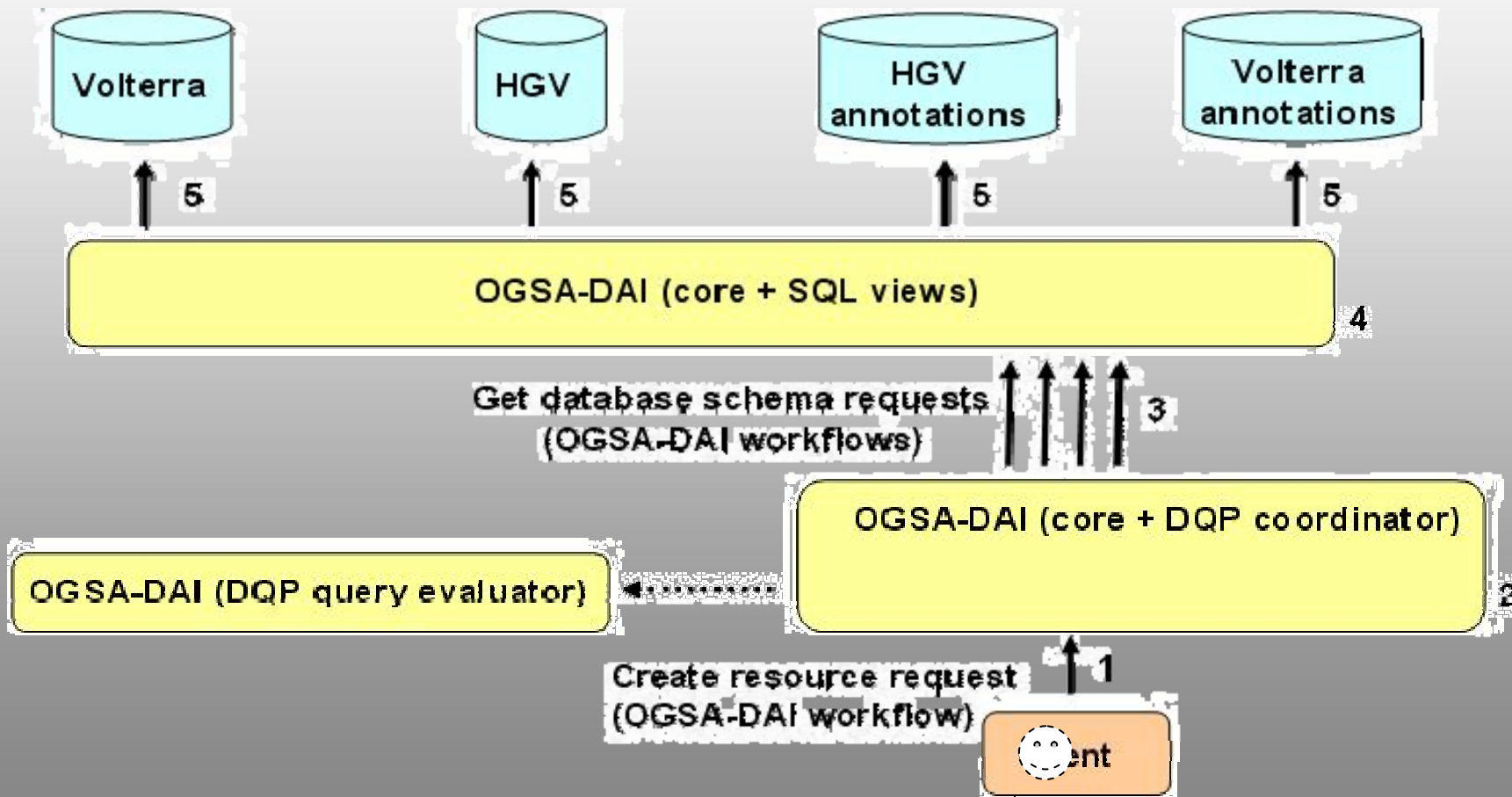
LAQVAT

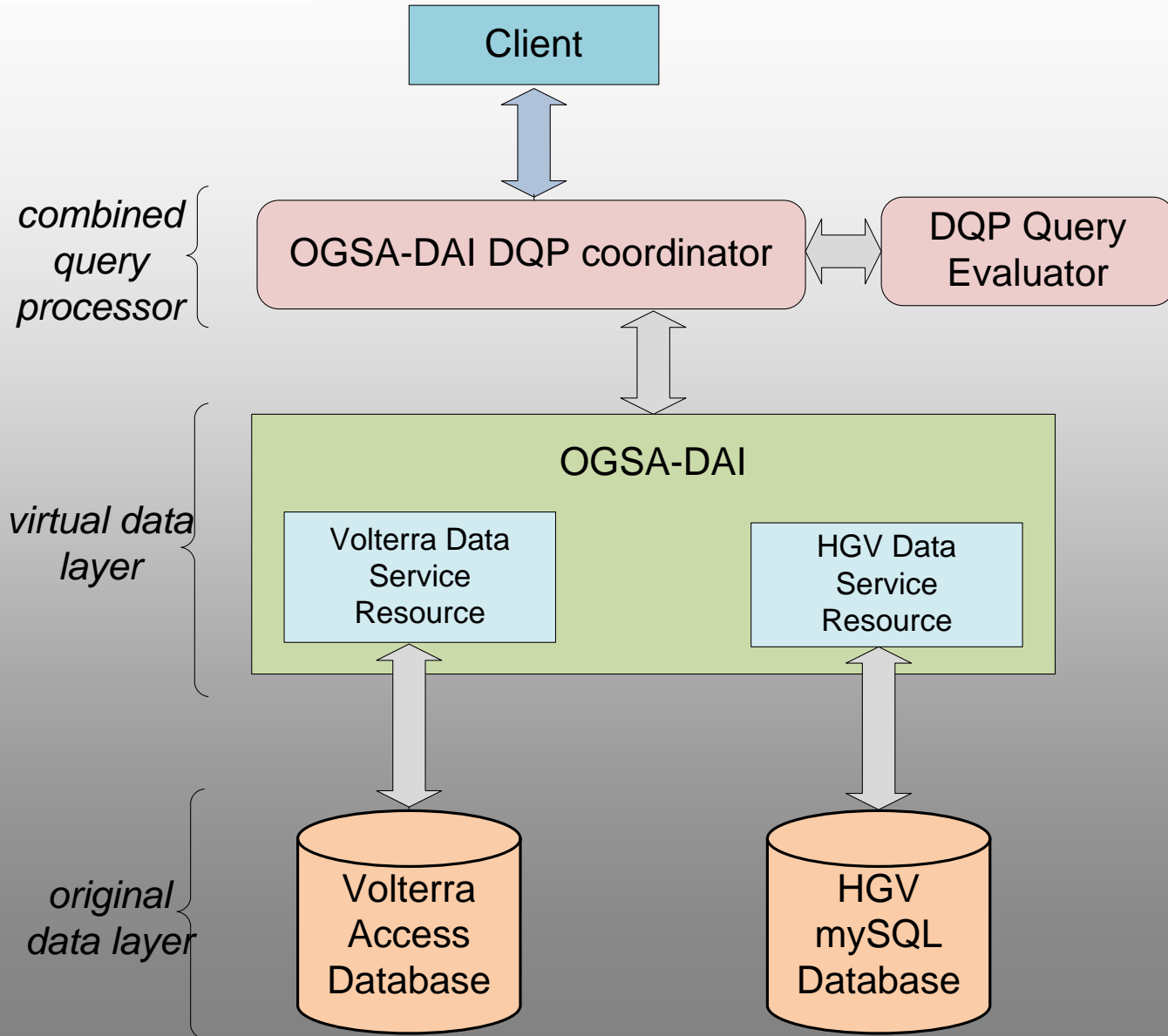
Lessons learned

Issues

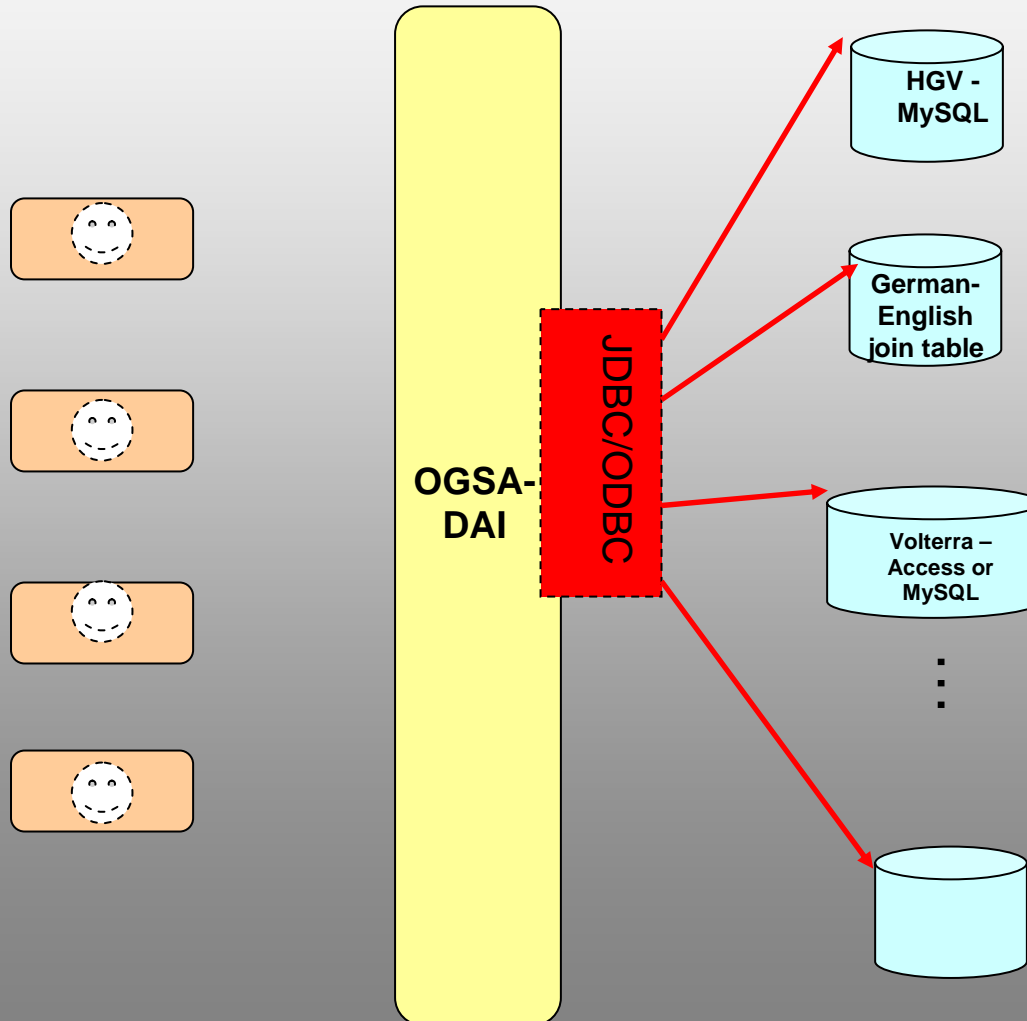
- Queries that really exploit joined-up structured data
- Drivers
 - Migrate the databases into something that can be used
- Problems with the technology

Architecture

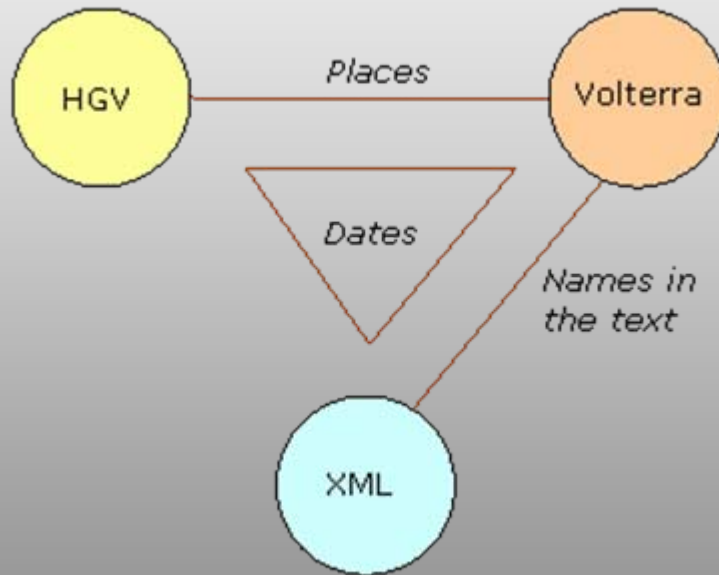




Vision: Virtual Data Centre



Queries



Possible links to the datasets

- 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

| A.Place | B.Time |
|----------------|---------------|
| Antiochia | 150 |
| Antiochia | 100 |
| Sitia | 200 |

- 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

JISC



engage
Engaging research with
e-Infrastructure

LAQVAT

Future Work

DARIAH

DARIAH

Architecture Overview

Feature - DARIAH: finding love on a grid

In the early 17th century, at the beginning of the Golden Age, elaborately illustrated secular and religious books began appearing on the Dutch market. They were a hit.

Was it the quality artwork? Did people read them for the articles? Or did their popularity stem from their subject matter, everyone's favorite topic: love.

With intriguing images and provocative text, the books helped young readers quickly learn about the many aspects of romance: choosing a partner, marital fidelity and the possible pangs of love.

Must modern readers miss out on such wisdom? No!

A digital database powered by grids

The [Emblem Project Utrecht](#), a digital humanities project out of the University of Utrecht in the Netherlands, is digitizing these books so they can be accessed via the Web.

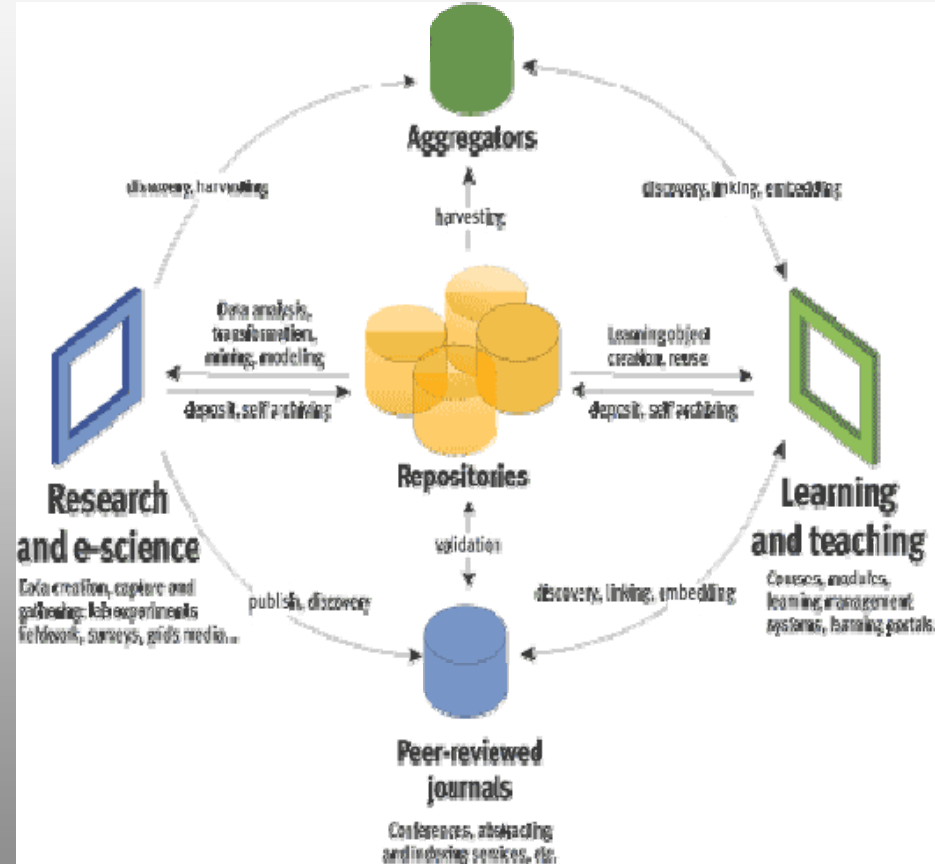


Love on the wing: digitized pages from 27 Dutch love emblem books are now accessible via the Web.
 Image courtesy of the Emblem Project Utrecht

Data-centric Collaboration

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

Dave de Roure: New e-Science Keynote



Network effects through data integration



DARIAH

Network of
Centres

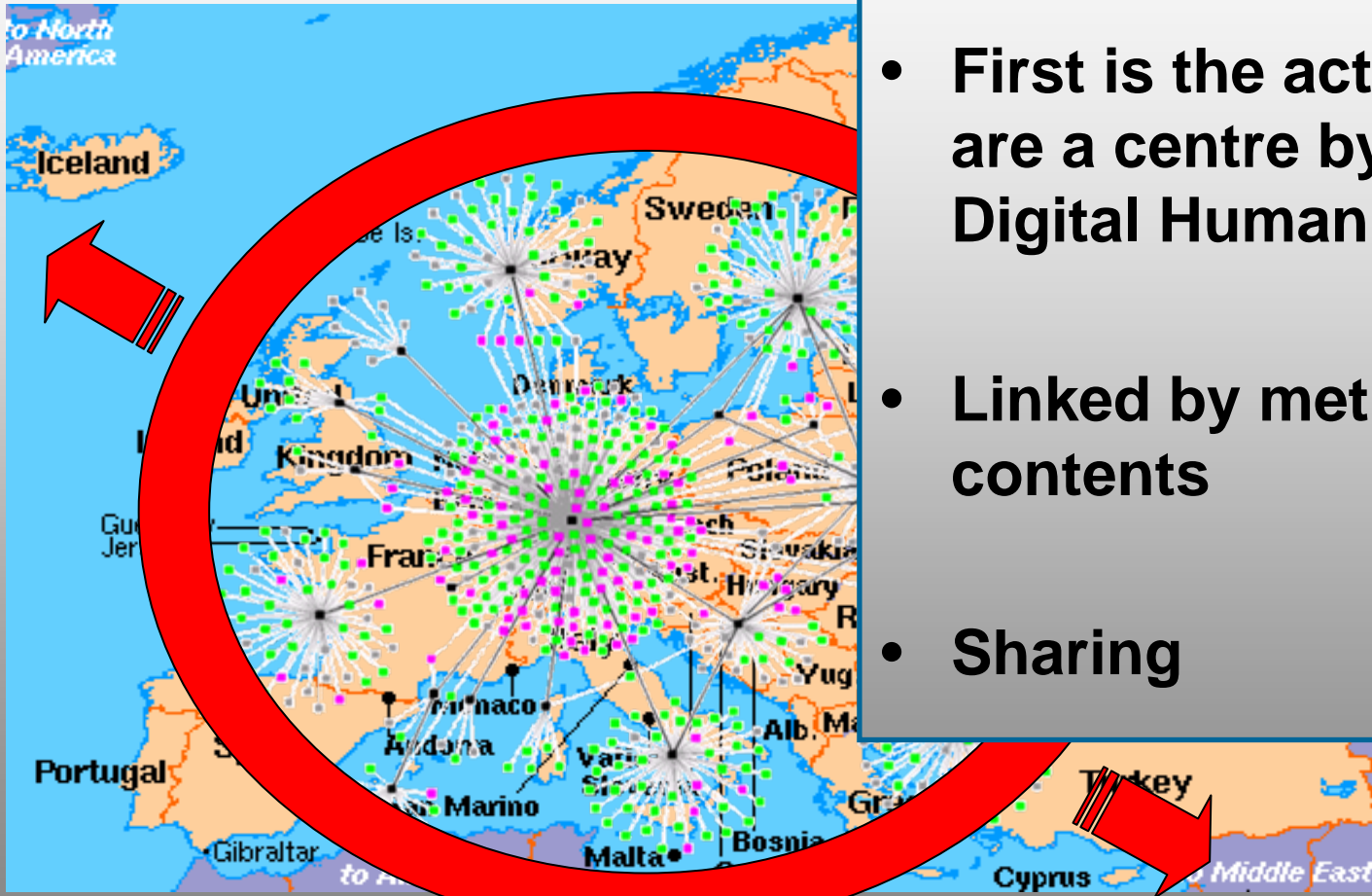
Methodological
Commons

Collective
Intelligence

Architecture
of Participation



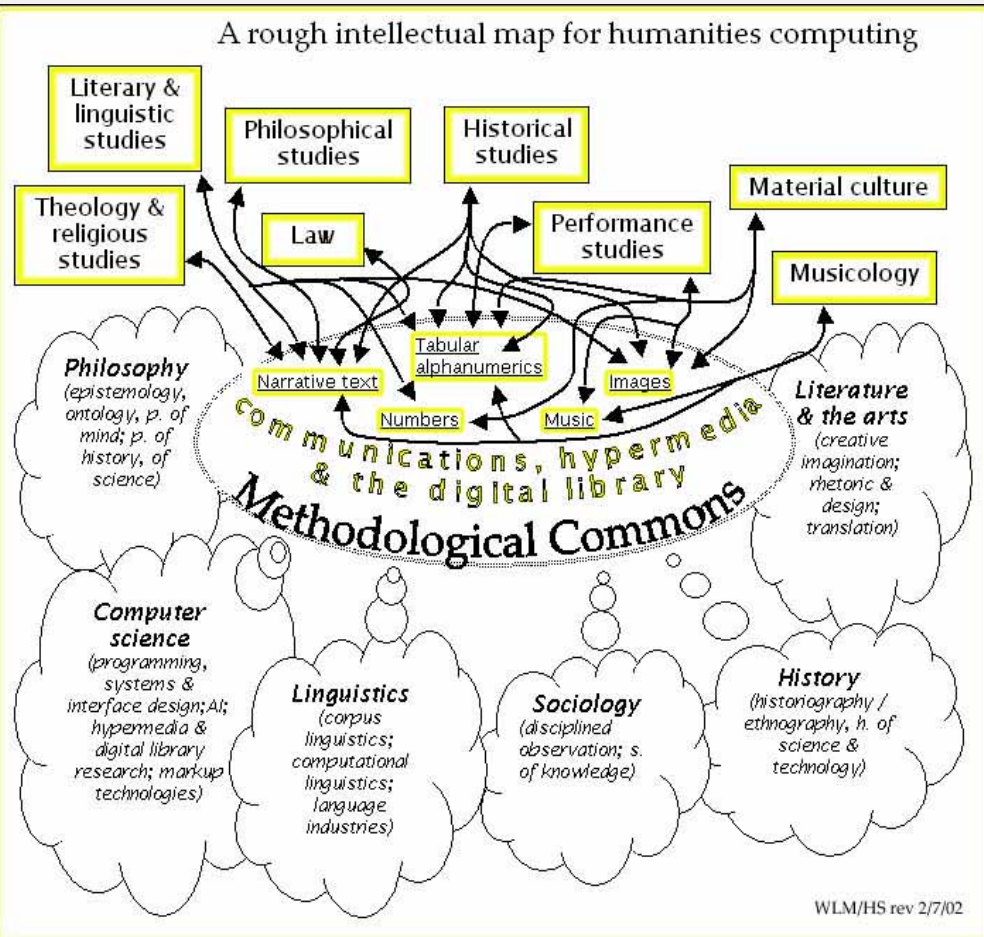
Network of Centres



- **First is the action: You are a centre by doing Digital Humanities**
- **Linked by methods and contents**
- **Sharing**

Methodological Commons

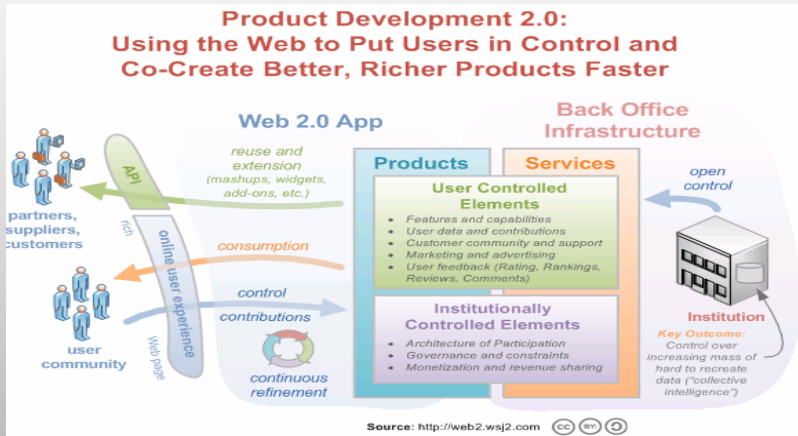
A rough intellectual map for humanities computing



WLM/HS rev 2/7/02

- **Methodological commons** of techniques applicable across disciplines. **Depend largely on the resources** employed rather than the subject
- **‘Scholarly primitives’** = ‘some basic functions common to scholarly activity’ (Unsworth)

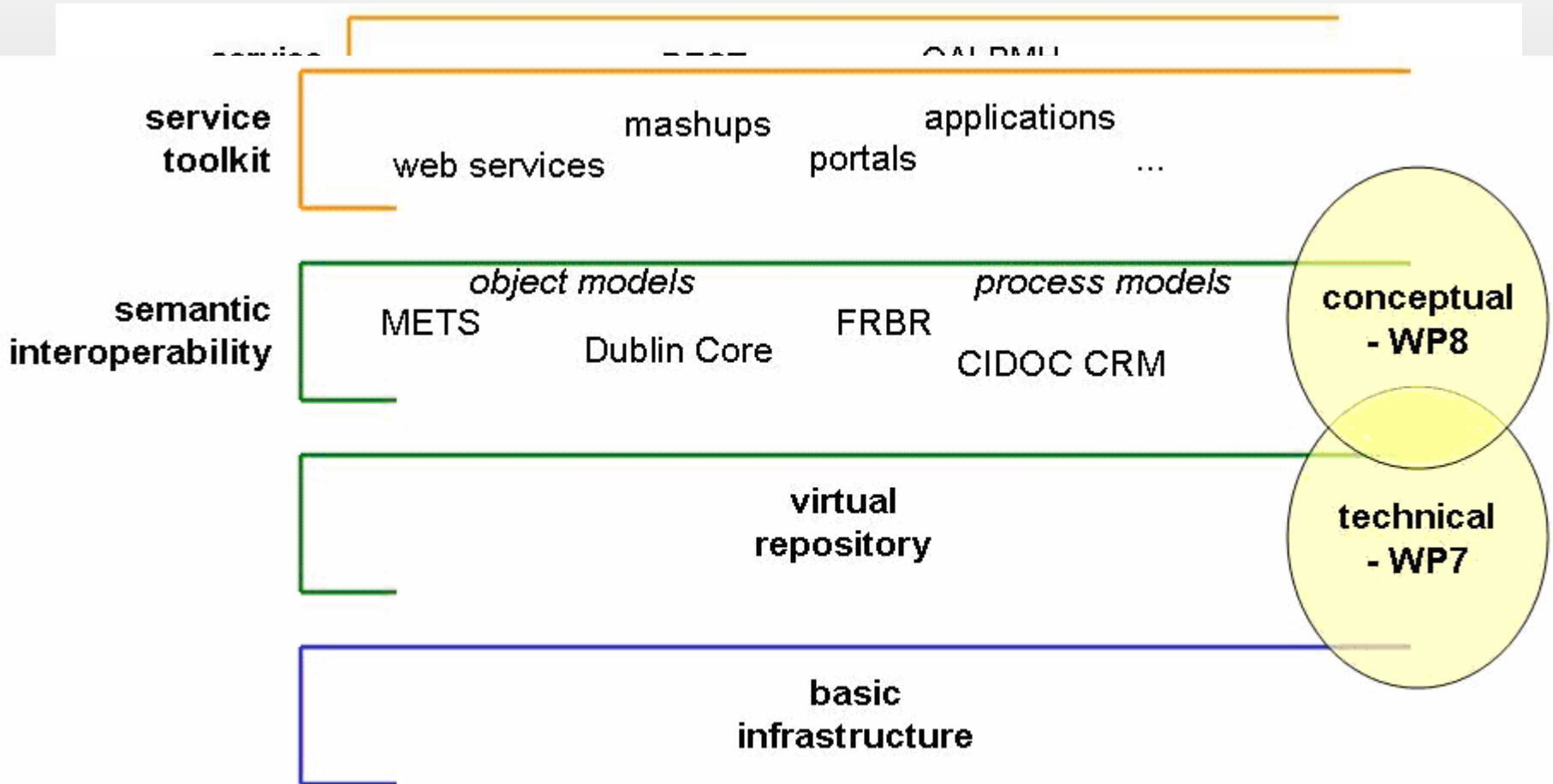
Collective Intelligence Architecture of Participation



- The better of Web 2.0
 - Openness
 - Peering: Redefined Authority
 - Sharing



Technical Architecture



Technology work: Solution and Semantics Focused

- **Two large demonstrators**
 - Integrating legacy applications
 - Building an exemplary solution for a node in the network
- **Various smaller experiments**
 - Data Backbone: Sharing of repository resources
 - Scholarly Applications: Tools and services for particular research
 - Service Genres: Search, etc.

Map Control:

Jump to:

Information:

Use the map on the right to choose a location



countries by
)

- E
- A
- A
- ap
- D

Warning! Searches are limited to 5,000 hits. If your search exceeds this limit use the where option to narrow your search.

Note: Due to firewall restrictions, University of York users will not be able to perform searches using Internet Explorer. While on campus, please use an alternative browser such as Mozilla Firefox or Netscape Navigator.

1. When

You can choose to search for site within a certain period range.

2. What

Choose a type of monument/site from a list of archaeological themes.

3. Where

Using a map of Europe choose the area you would like to search.



TEI demonstrator



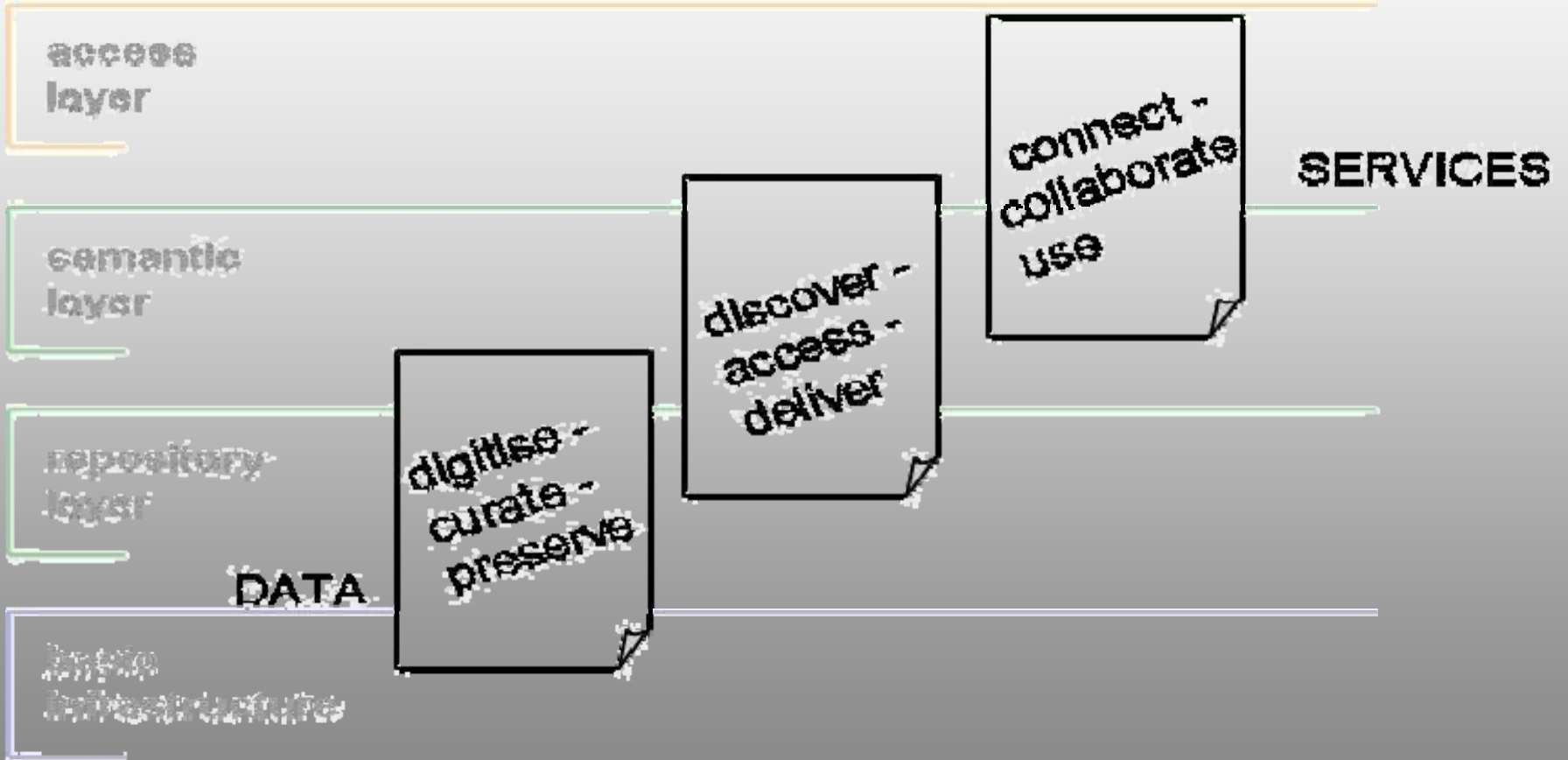
NORDISK FORSKNING SINSTITUT
KØBENHAVNS UNIVERSITET

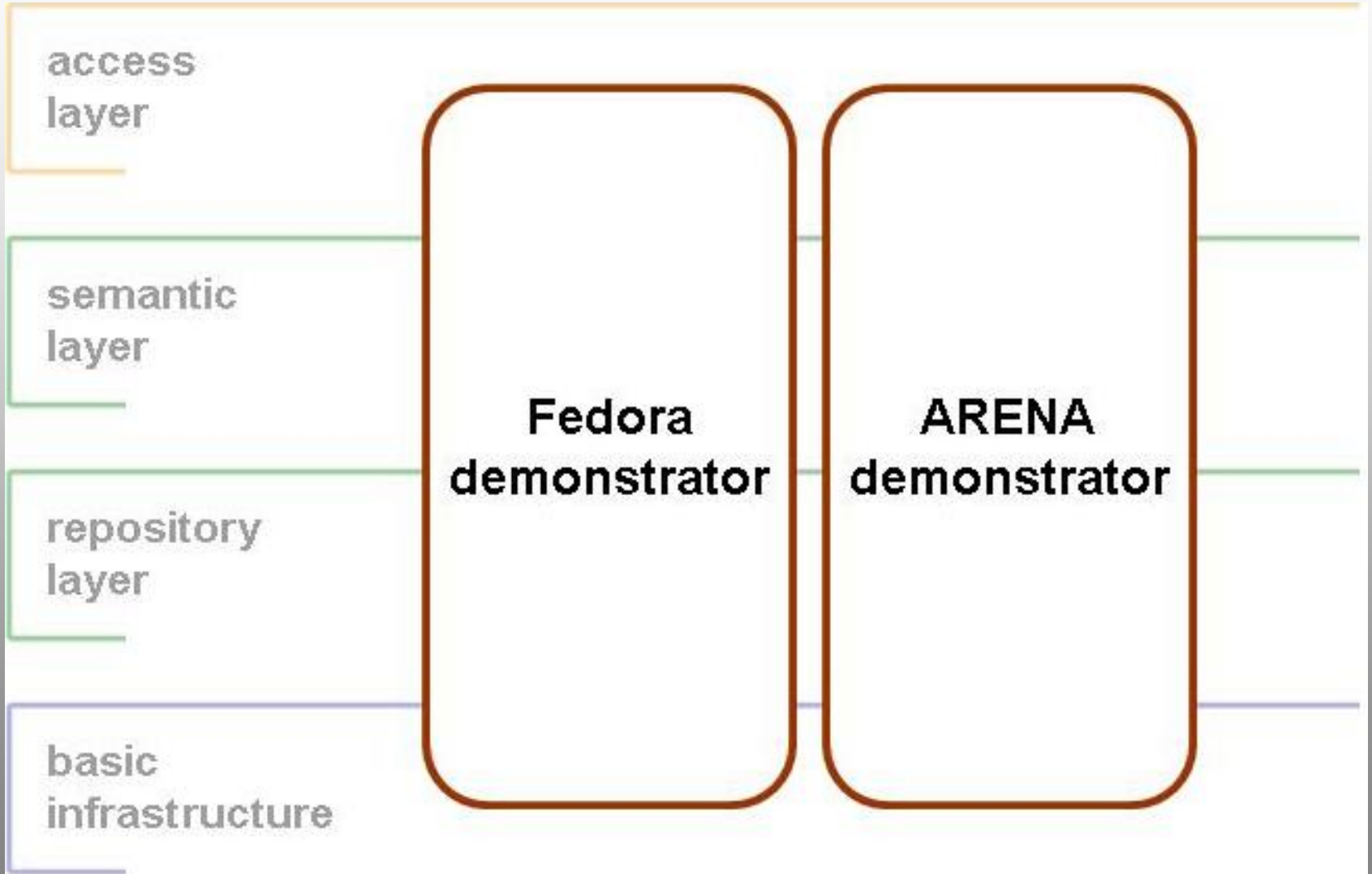


- Exemplary digital archive
- Collections at the Nordisk Forskningsinstitut (NFI) as well as reference works
- Fedora and Grid integration (flat files)



DARIAH Concept







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

JISC



engage
Engaging research with
e-Infrastructure

LAQVAT

Questions

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