

Grid enabling data sets with OGSA-DAI

Ally Hume

e-Infrastructures in the Arts and Humanities and Social Sciences:
Grid-enabling Data Sets

5th December 2007



omiieurope
open middleware infrastructure institute



National
Science
Centre

epcc



OGSA-DAI

the globus alliance
www.globus.org



EPSRC

Engineering and Physical Sciences
Research Council



omii-uk

Web: www.omii.ac.uk

Email: info@omii.ac.uk

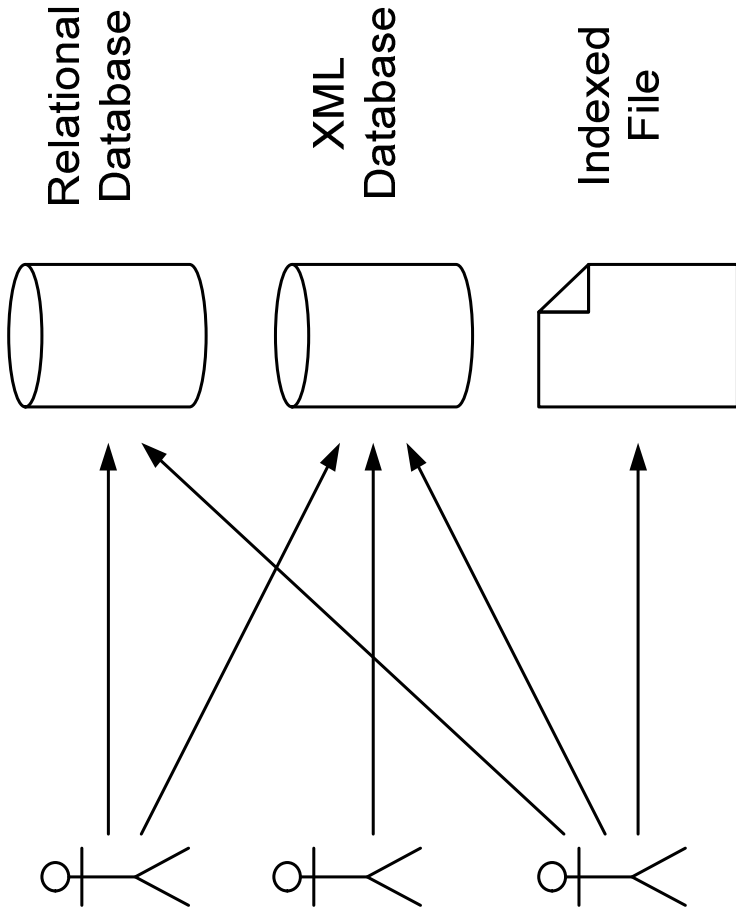
What is OGSA-DAI?

- An **extensible** framework
- accessed via **web services**
- that executes **data-centric workflows**
- involving **heterogeneous** data resources
- for the purposes of **data access, integration, transformation** and **delivery**
- within a **grid**
- and is intended as a **toolkit** for building higher-level **application-specific** data services

Sharing data in a grid

Motivation

- Grid is about sharing resources
- OGSA-DAI is about sharing structured data resources



Sharing data via web site download

- ZIP up data and put it on a web site
- Pros
 - Easy distribution for providers
 - Easy access for consumers
- Cons
 - Consumers have to download all the data
 - Consumers have to load data into local databases to use it
 - Static snapshot
 - Security

Sharing data via direct access

- **Providers tell consumers**
 - Database URL – `mycomputer.epcc.ed.ac.uk:3306`
 - Username – `userID`
 - Password – `password`
- **Pros**
 - Consumers have direct access
- **Cons**
 - Firewall issues
 - User and password management is hard
 - No consistent security model
 - Hard to use in grid/web service workflows

Sharing data via direct access

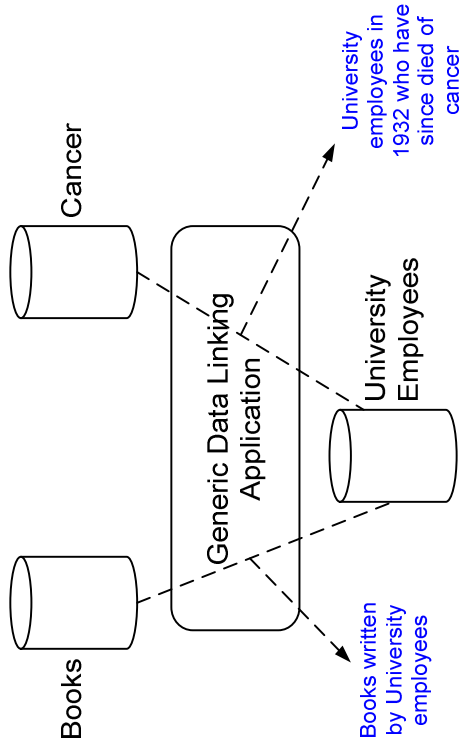
- **Cons (continued)**
 - No server-side layer in which to standardize database heterogeneities
 - Myriad drivers
 - Different APIs across different data types
 - Relational and JDBC
 - XML and XMLDB
 - Indexed files and Lucene

Domain-specific web services

- Manipulate data using domain-specific operations, e.g.
 - Book findByISBN(ISBN)
 - List<Book> findByAuthor(Author)
 - List<Book> findByKeyword(Word)
- Pros
 - Fits with grid/web service approach
 - Abstraction hides back-end database details
 - Web services are programming language neutral
 - Operations likely to map well to authorization policies

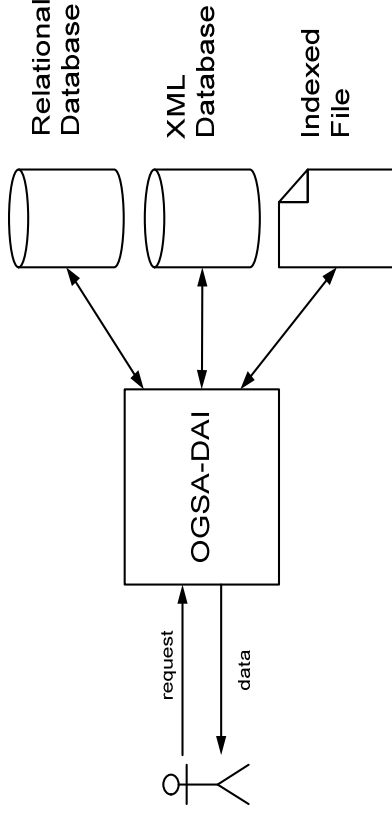
Domain-specific web services

- Cons
 - Slower than direct access
 - Web service layer
 - SOAP transport overhead – especially for large result sets
 - Domain-specific API prevents use of generic data exploration, mining and manipulation tools

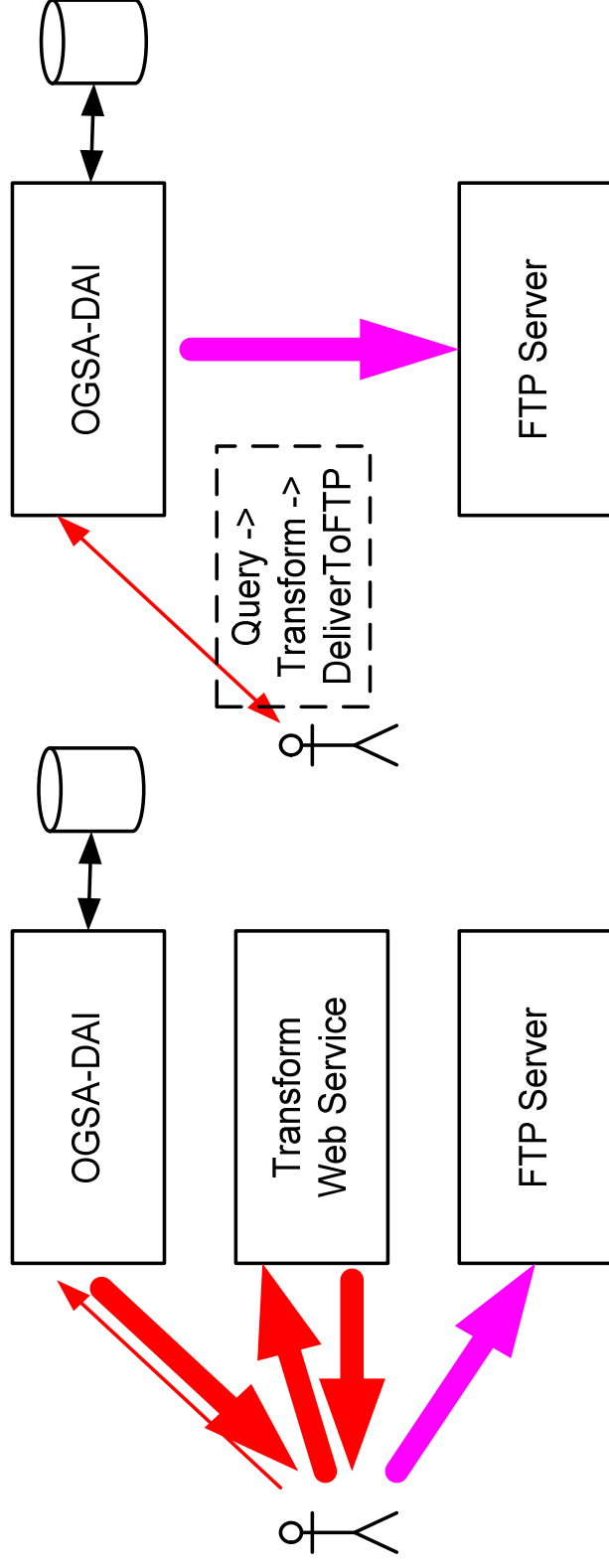


OGSA-DAI generic web services

- Manipulate data using OGSA-DAI's generic web services
- Clients sees the data in its 'raw' format, e.g.
 - Tables, columns, rows for relational data
 - Collections, elements etc. for XML data
- Clients can obtain the schema of the data
- Clients send queries in appropriate query language, e.g. SQL, XPath



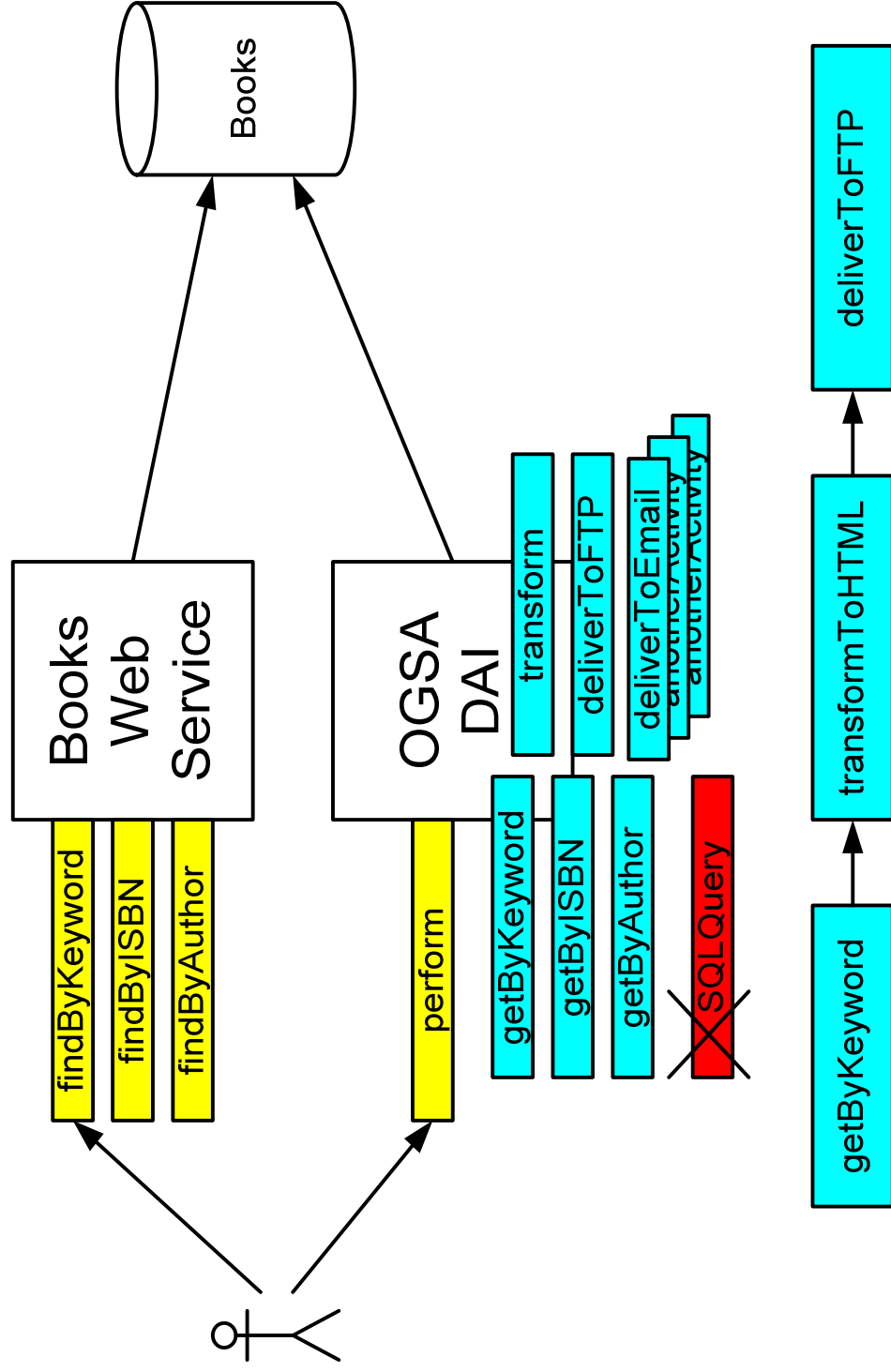
Getting away from SOAP – workflows



Getting away from SOAP

- Asides from FTP there is also...
- SOAP attachments
 - Data comes along with, but external to, a SOAP message
- GridFTP
- E-mail
- ...

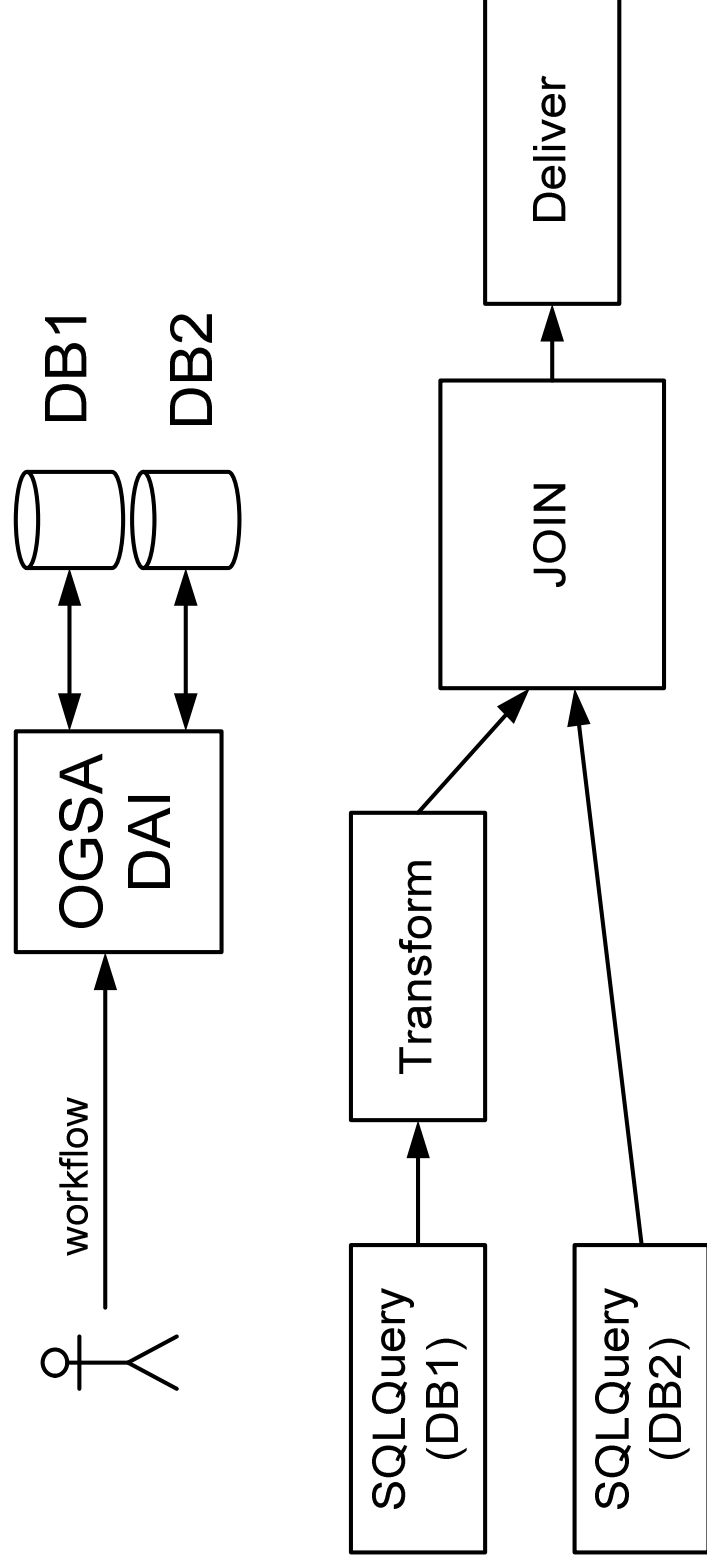
Example – application-specific activities



Data integration examples

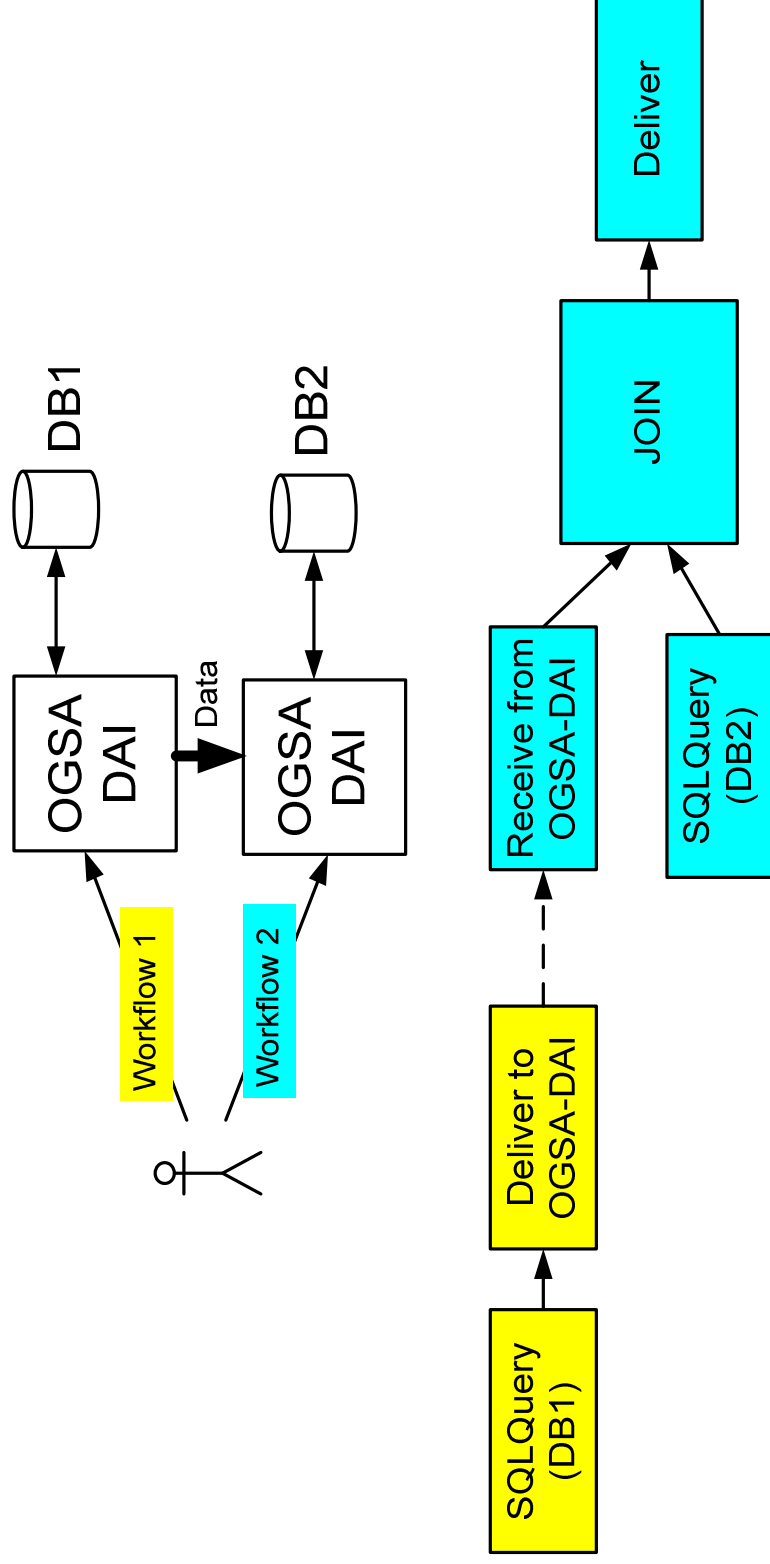
Data integration with OGSA-DAI workflows

- Using a single workflow

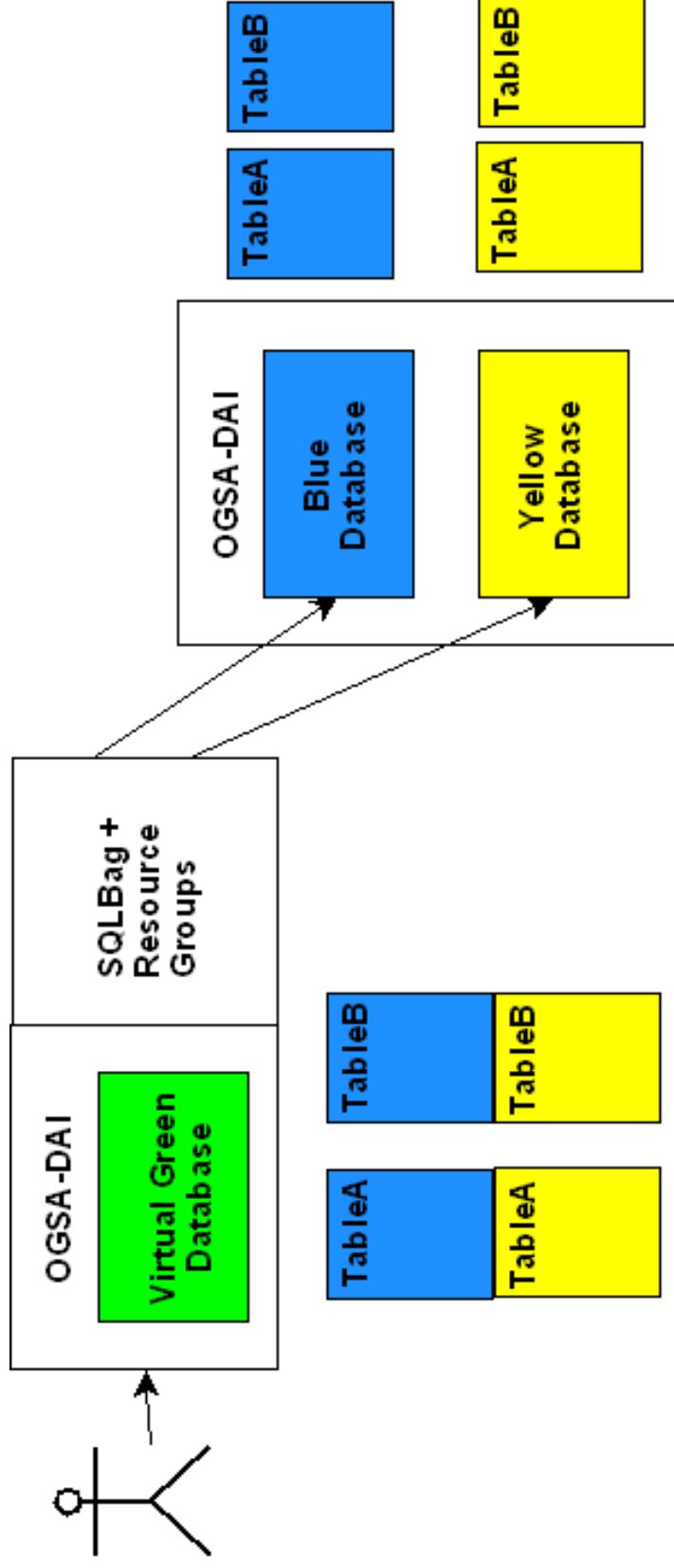


Data integration with OGSA-DAI workflows

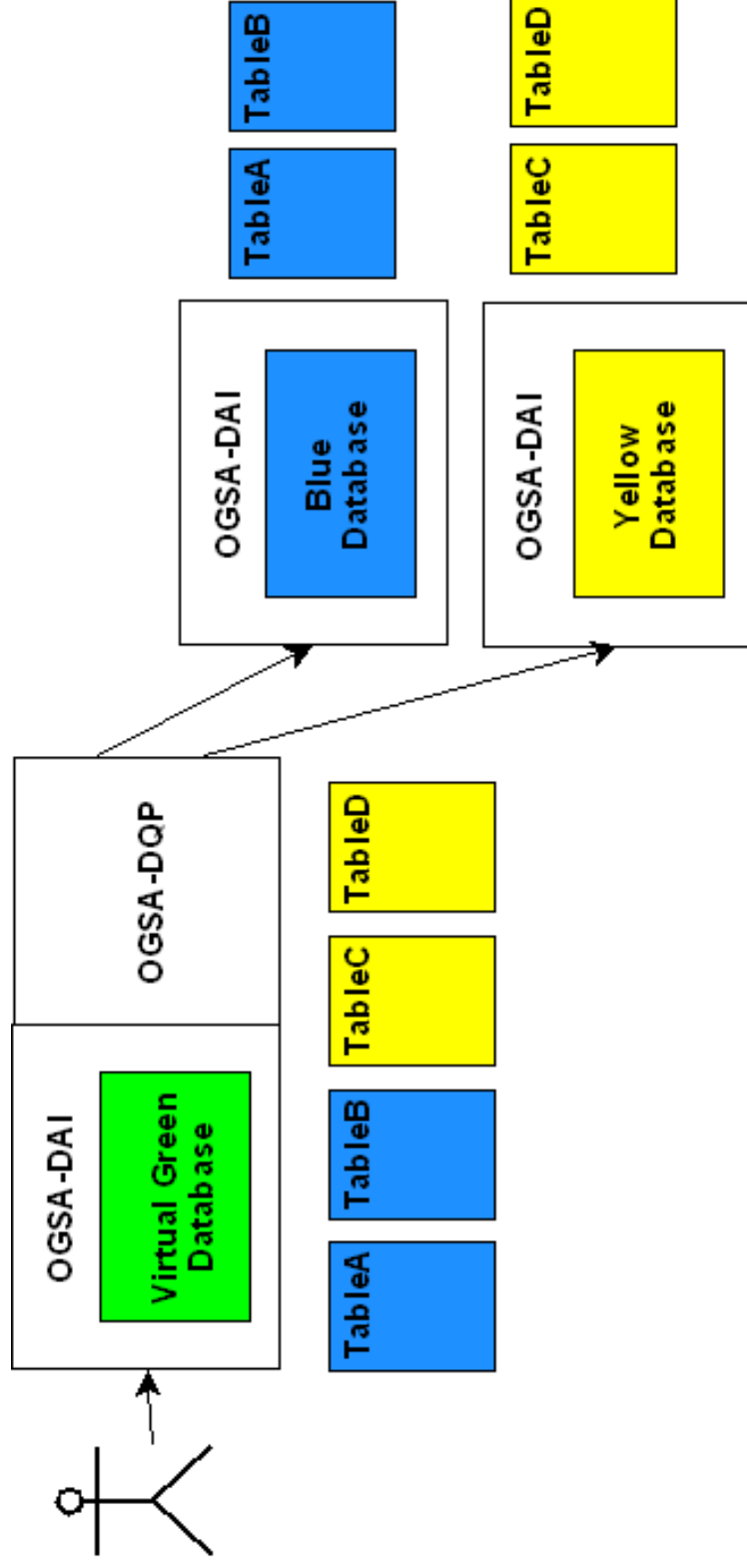
- Across OGSA-DAI services



Vertical hidden data federation



Hidden horizontal data federation



View and transforms

- Schema mapping using views and transforms
 - boilingPoint: `SELECT chemical, tempInKelvin - 273 AS tempInCelsius FROM bPoints`
 - trees: `SELECT DATE(SplitTrees.Date), SplitTrees.Site, SplitTrees.Species, SplitTrees.Density FROM SplitTrees.Func(Trees) AS SplitTrees(Date: VARCHAR, Site: VARCHAR, Species: VARCHAR, Density: FLOAT)`
 - `SELECT name, age from patients where doctor = $DN`

GEODE example

GEODE

- Grid Enabled Occupational Data Environment
- Repository for occupational data
 - e.g. gender segregation for occupational codes used in UK 1991 census (SOC90)
- Functionality to link researcher's data with these data sets.
- See: <http://www.geode.stir.ac.uk/>



Web: www.omii.ac.uk

Email: info@omii.ac.uk

GEODE cont

- Login to portal
- Search for 'gender'
- Discover "Gender segregation statistics from Hakim 1998, for UK SOC90, using 1991 census".
- Decide to link this to my local data
- Java Web Start application downloads...

Grid Enabled Occupational Data Environment

Portlet for searching data resources (G1 and deposited)

Resource ID	Resource Title	Abstract	Supplier	Supplied Date	Original Creator	Original Publication Date	Affiliation	Email	Occupational Classifications Used
61	Gender segregation values for SOC-90 data (Hakim 1998)	View Abstract	Paul Lambert	2003-04-09	Catherine Hakim	1998	University of Strirling	paul.lambert@string.ac.uk	SOC-90 (3-digit) (Standard Occupational Classification the Labour 1995)
63	AWM for sweden	View Abstract	Erik Bihagen	2006-11-22			Stockholm University	erik.bihagen@sofi.su.se	
102	Gender segregation tables UK	View Abstract	Paul Lambert	2007-06-09	Equal Opportunities Commission	2005	University of Strirling	paul.lambert@string.ac.uk	Standard occupational classification 2000 major groups (SOC2000 1-digit)
114	Gender segregation statistics for ISCO-88, from Hakim (1998)	View Abstract	Paul Lambert	2007-06-13	Catherine Hakim; Paul Lambert	1998	University of Strirling	paul.lambert@string.ac.uk	ISCO88 minor groups (Standard Occupational Classification of Occupations; 3-digit minor groups)

Results of index search

Title	Authors	Countries	Time Periods	Service URI	Resource ID	Corresponding Uncurated Resources	View DDI
Gender segregation statistics for UK SOC90, using 1991 census	Paul Lambert	United Kingdom	1990	http://139.153.254.158:8686/wsrf/services/geode/GEODFactory/Service/paulambert-hakimsoe61			View DDI

Hit Enter or click Find Resources Help on search Search: gender

GEODE cont.

Occupational Matching Client v0.1

Path of your data (CSV)

Have DDI? No Yes Path of your DDI

Preview of your selected data

name	soc90	status	something
A. Judge	240	7	100
A. Nurse	340	7	200
A. Geek	214	7	300

Please select the matching variables

M/O	Category URI	Var matching specified category	Default Value
M	http://www.geode.stir.ac.uk/bugs.html#soc90	soc90	0000

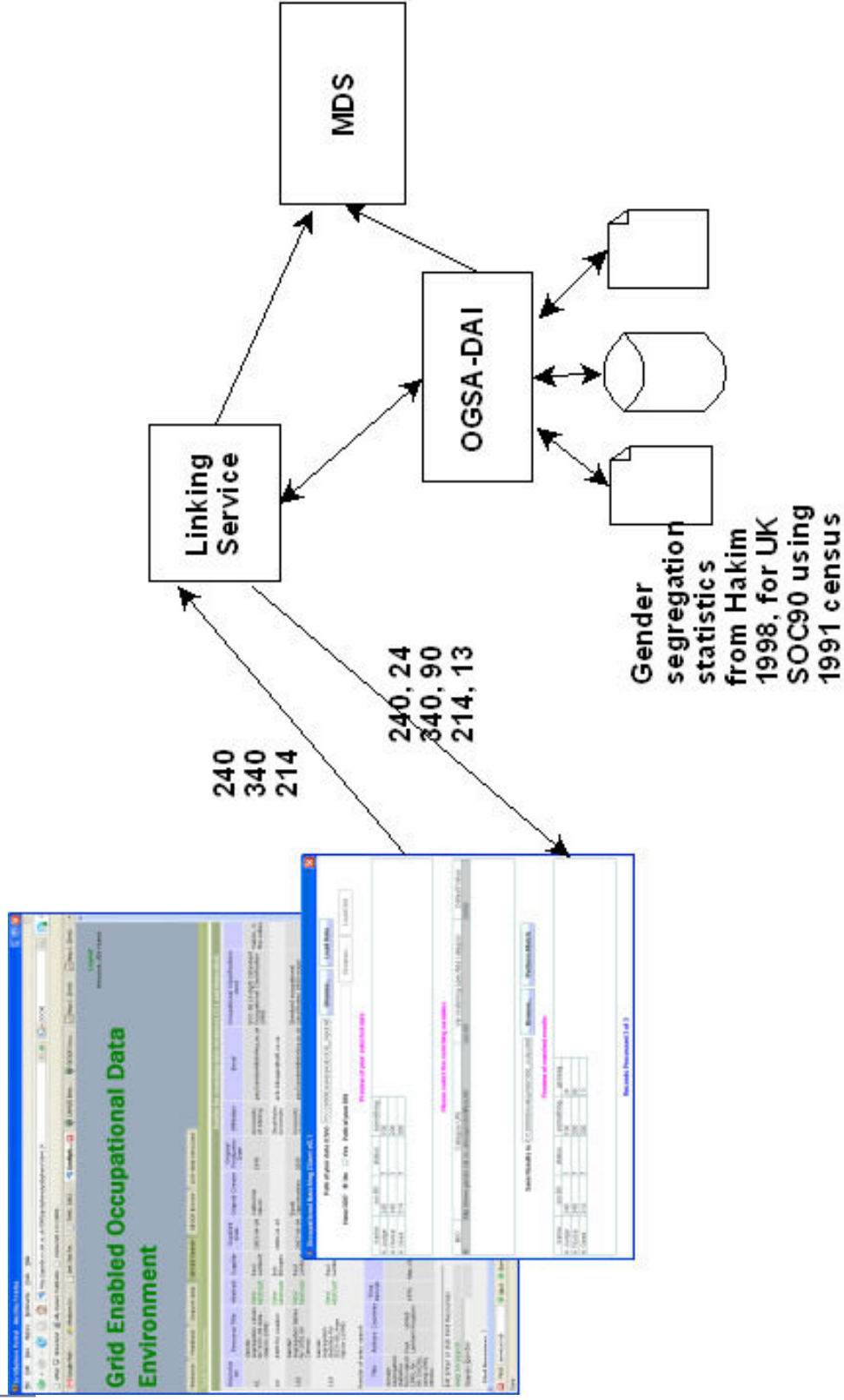
Save Results in

Preview of matched results

name	soc90	status	something	gensseg
A. Judge	240	7	100	24
A. Nurse	340	7	200	90
A. Geek	214	7	300	13

Records Processed 3 of 3

GEODE cont.



SEE-GEO example

SEE-GEO

- SEcurE access to GEOspatial services
 - EDINA, NeSC, NCESS, MIMAS
 - <http://edina.ac.uk/projects/seesaw/seegeo/index.html>



- Access to geospatial information on a grid
- Open Geospatial Consortium (OGC) web services

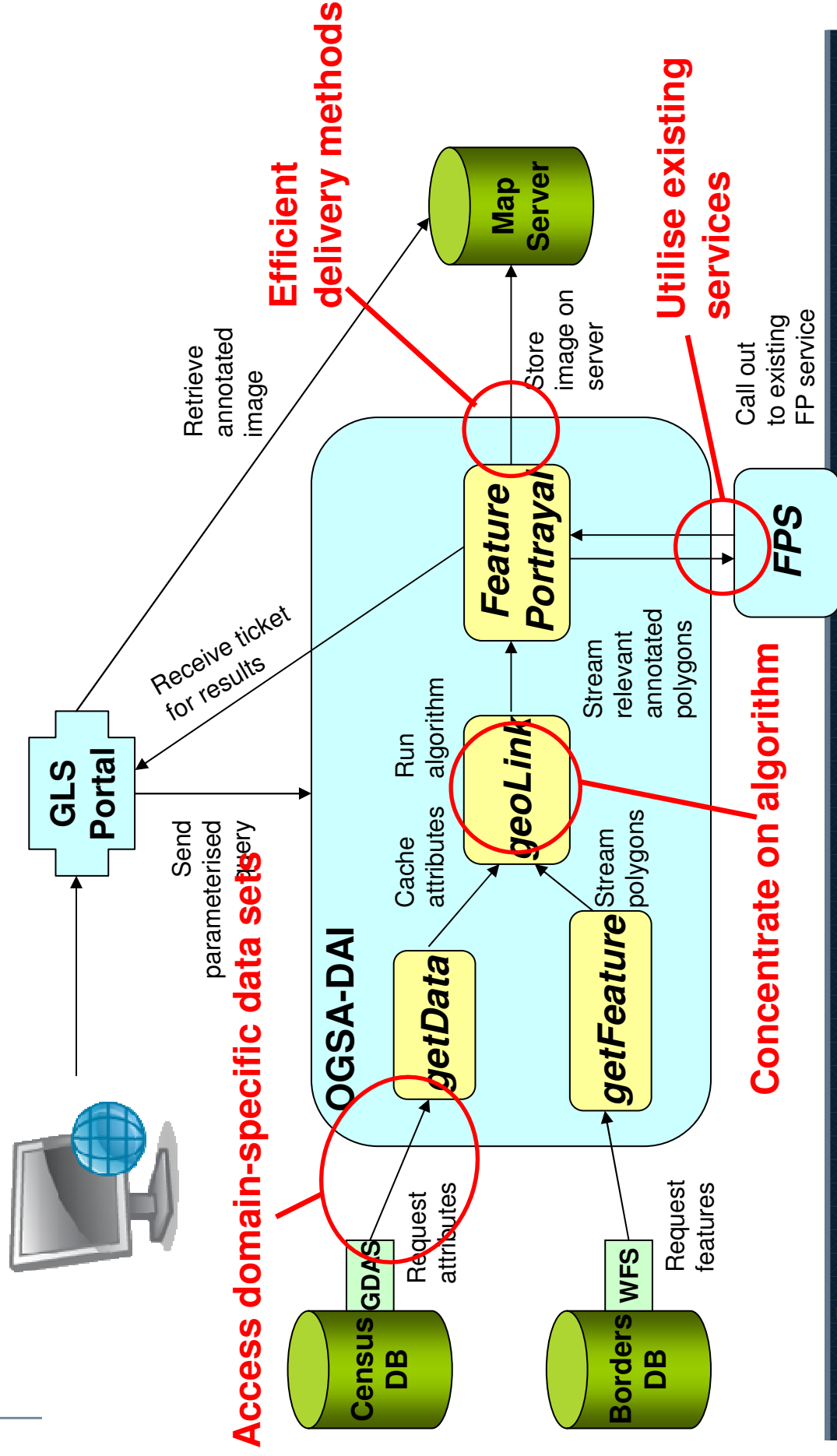
SEE-GEO cont.

- OGSA-DAI being extended to offer integrated, distributed resource management for geo-spatial tools
- Using established open interoperability standards
- Web Feature Service (WFS) and Web Map Service (WMS) integrated into OGSA-DAI
- The IE is hardening candidate OGC specifications
 - Geolinked Data Access Service (GDAS)
 - Geolinking Service (GLS)
- Validate
- Web Coverage Service (WCS) scheduled
- Extend to support secure access

SEE-GEO demonstrator

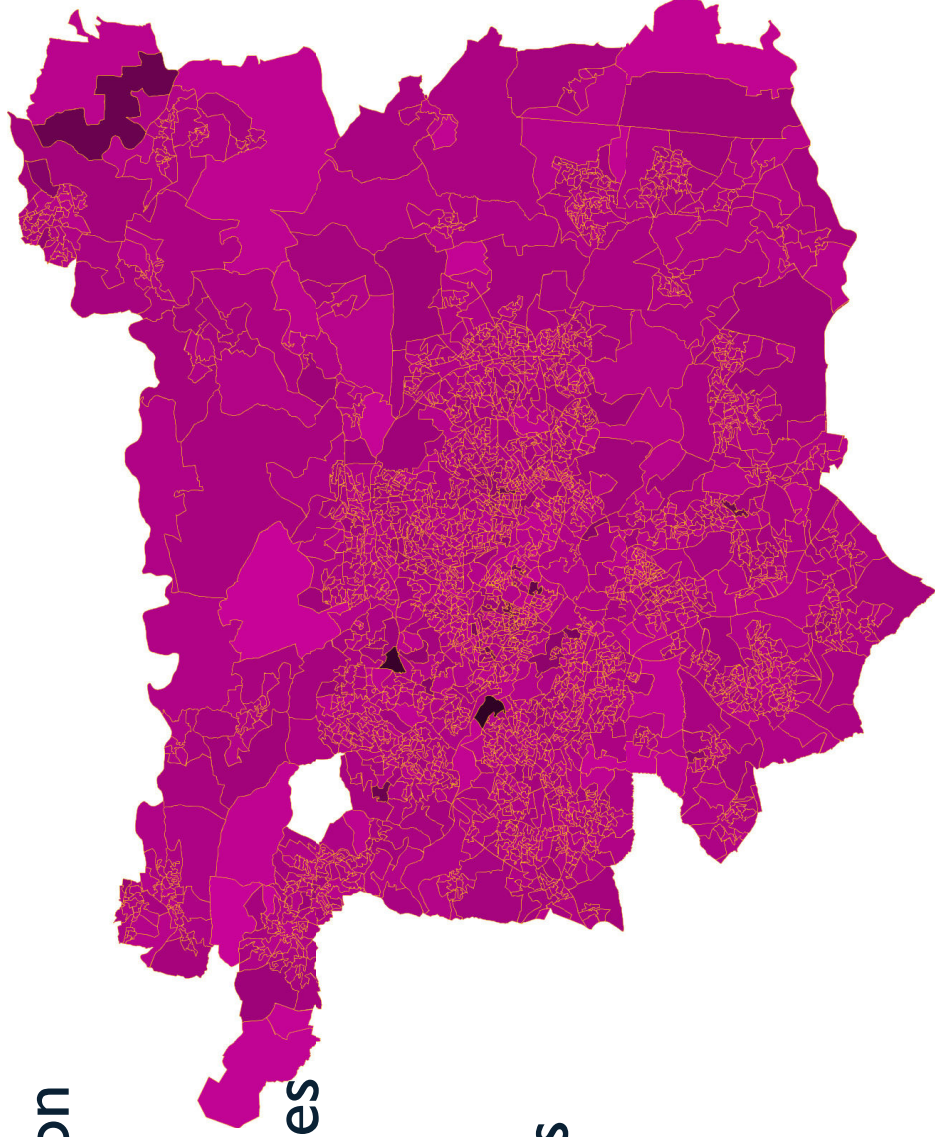
- Two data resources
 - Census statistics
 - Attributes about a region
 - e.g. population
 - Geo-data access service (GDAS)
 - Borders data
 - Unique regions encoded as polygons
 - Web feature service (WFS)
- How to link the attributes to the regions?
- A geo-linking service
 - Execute a join across the two data sets
- Implemented as a Web Processing Service (WPS)

Demographic forecasting

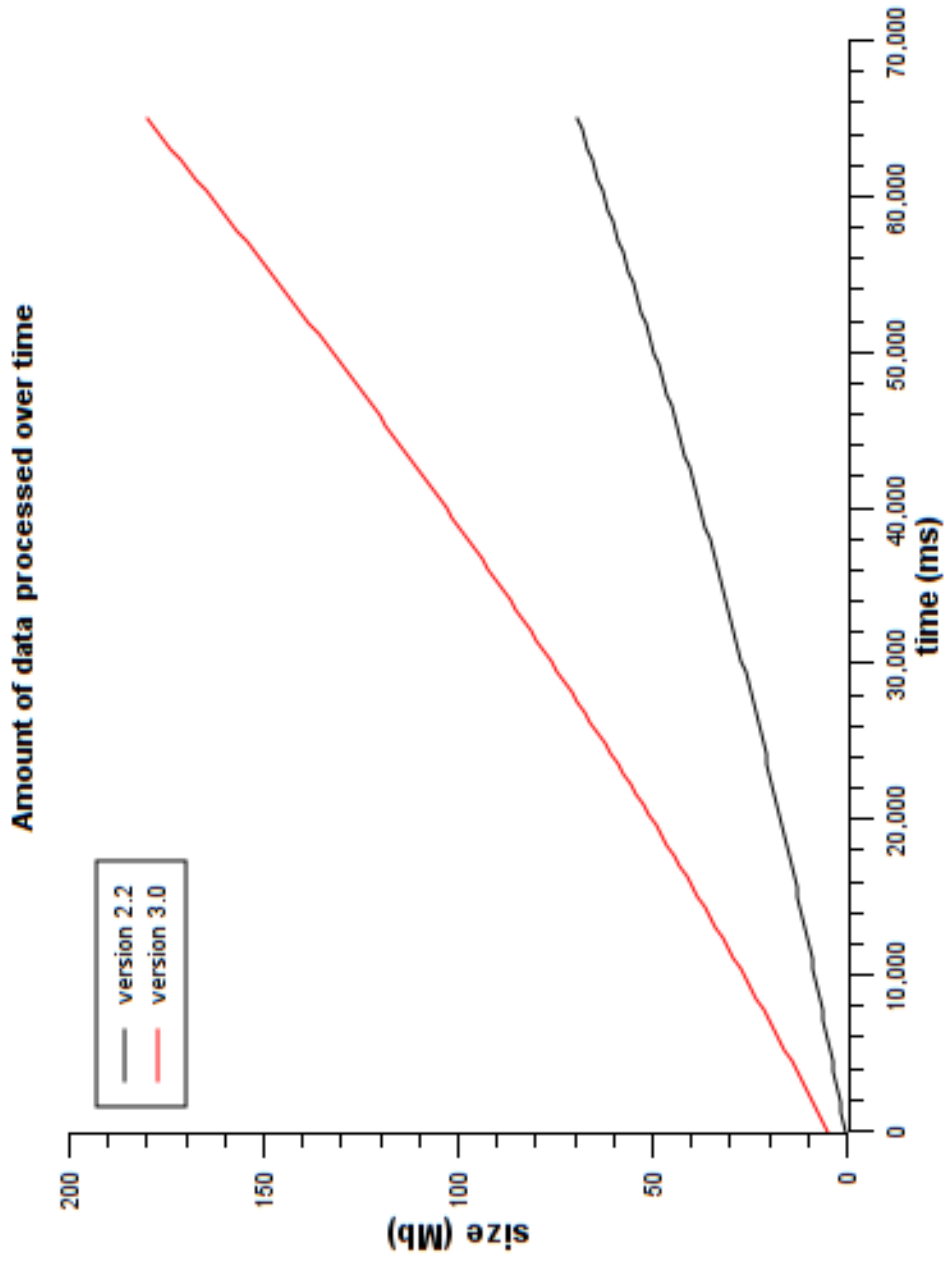


SEE-GEO output

- Leeds population distribution by census output area. Boundaries from EDINA, 2000 census statistics from MIMAS



SEE-GEO performance



What did OGSA-DAI give SEE-GEO?

- Could implement GLS service without OGSA-DAI
- But using OGSA-DAI allowed leverage of
 - Workflow engine
 - Out-of-the-box activities for
 - Queries
 - Delivery
 - Security
 - Other grid technologies e.g. GridFTP

What did OGSA-DAI give SEE-GEO?

- A toolkit to
 - Develop domain-specific activities
 - Develop support for domain-specific data resources
 - Ability to execute workflows using these
 - Build OGC Web Processing Services (WPS)
- Relatively little effort to
 - Choose different data resources dynamically
 - Merge GDAS XML into a relational data resource
 - Transfer data using GridFTP
 - Protect data using GSI
 - Experiment!

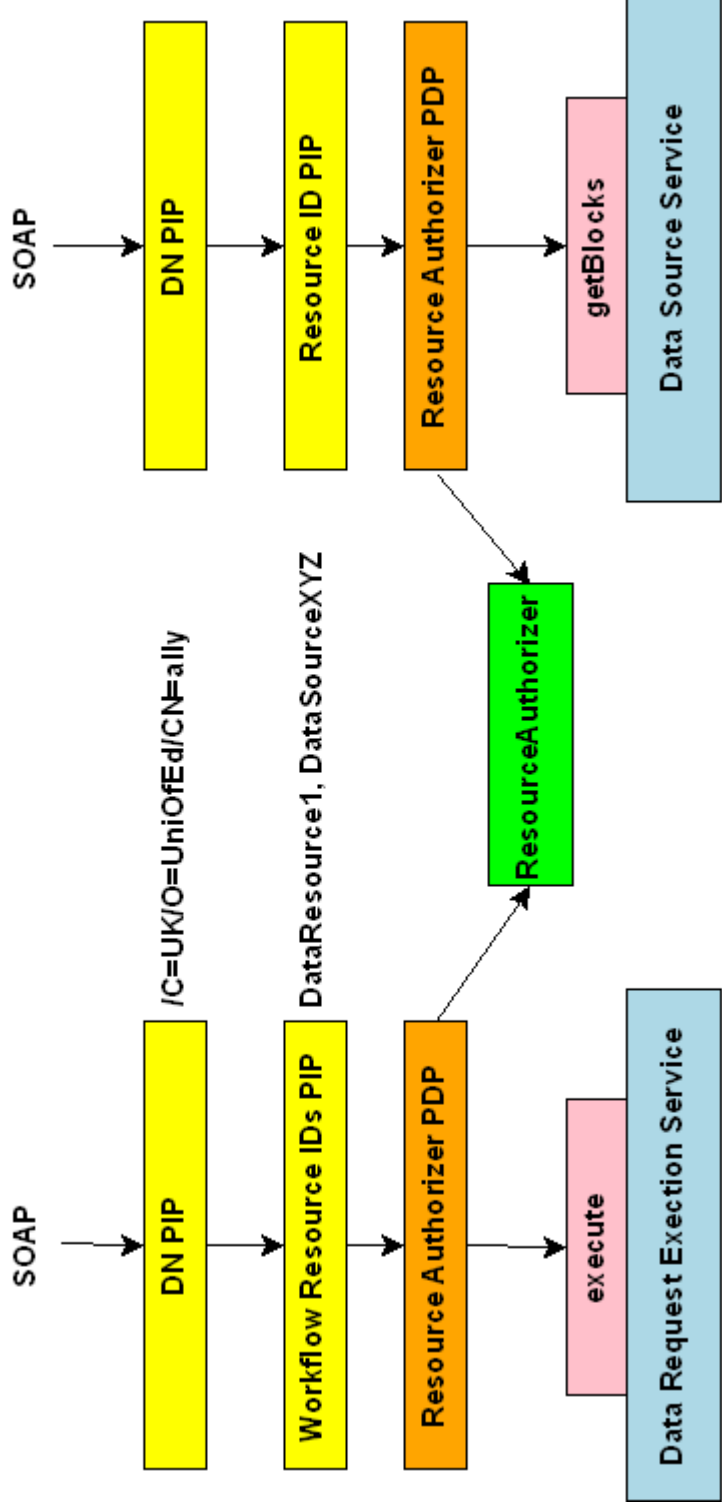
What next for SEE-GEO?

- Add Web Coverage Service (WCS)
- Look at security and OGC
 - Shibboleth
 - Grid Security Infrastructure (GSI)
 - Privilege and Role Management Infrastructure Standards Validation (PERMIS)
- OGSA-DAI:Z39.509 SRW/U bridge
 - Ordnance Survey Master Map delivery using a grid

OGSA-DAI and authorization

OGSA-DAI authorization

- Authorization on incoming SOAP request



OGSA-DAI authorization cont.

- **SecurityContext object**
 - One for each request
 - By default contains DN and credential
- **Login provider plug-in objects**
 - One for each relational resource
 - Maps SecurityContext -> database login
- **ResourceAuthorizer plug-in object**
 - Used by ResourceAuthorizerPDP
 - Listens to event from resource manager
- **RuntimeWorkflowAuthorizer plug-in object**
 - Authorizes dynamically created workflows at runtime