



SEMUSE Workshop
Semantic Web and CIDOC CRM 2009
Washington DC
25th October 2009

Semuse

The Future of the Past:
Using CIDOC CRM for CLAROS
(Classical Arts Research Online Services)



David Shotton

Image BioInformatics Research Group

Department of Zoology

University of Oxford, UK

<http://ibrg.zoo.ox.ac.uk>

e-mail: david.shotton@zoo.ox.ac.uk



Outline

- How we got started with data webs
- Principles of data webs
- OpenFlyData, a data web for *Drosophila* gene expression data
- CLAROS: Classical Arts Research Online Services
- Use of CIDOC CRM

2005: First ideas about integrating distributed data



- In 2005, we were developing the **BioImage Database**
- Original high resolution microscopic images could be stored in their original distributed locations, rather than in the BioImage Database with their metadata
- One day, I had the idea that, if you could fetch the images from distributed resources into the user's browser on the fly, **why not also fetch the metadata?**
- In other words, the database transforms into a **data web**

2006: My first presentation on data webs

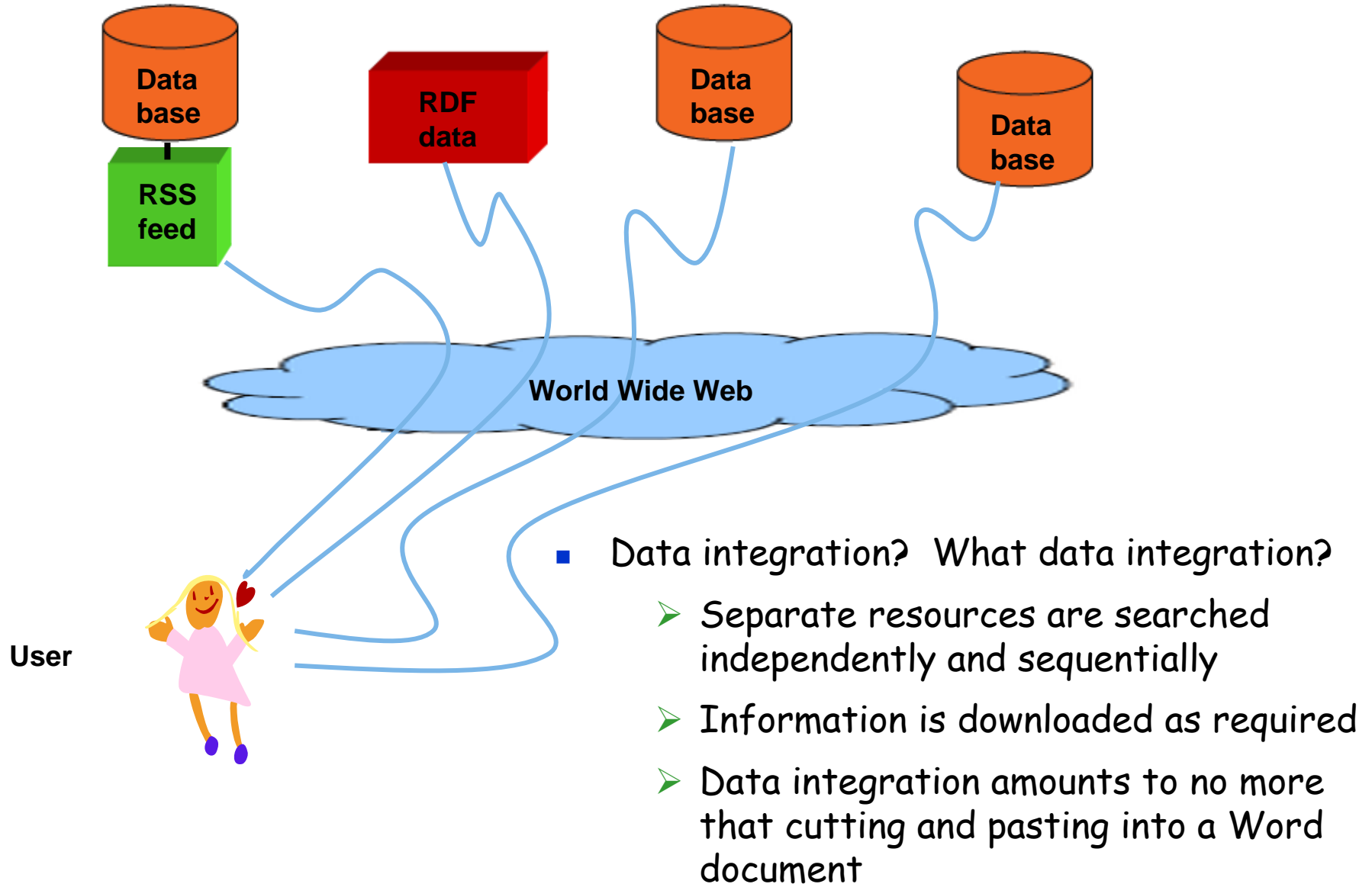
**Semantic Interoperability for e-Research
in the Sciences, Arts and Humanities**

Imperial College
March 30th 2006

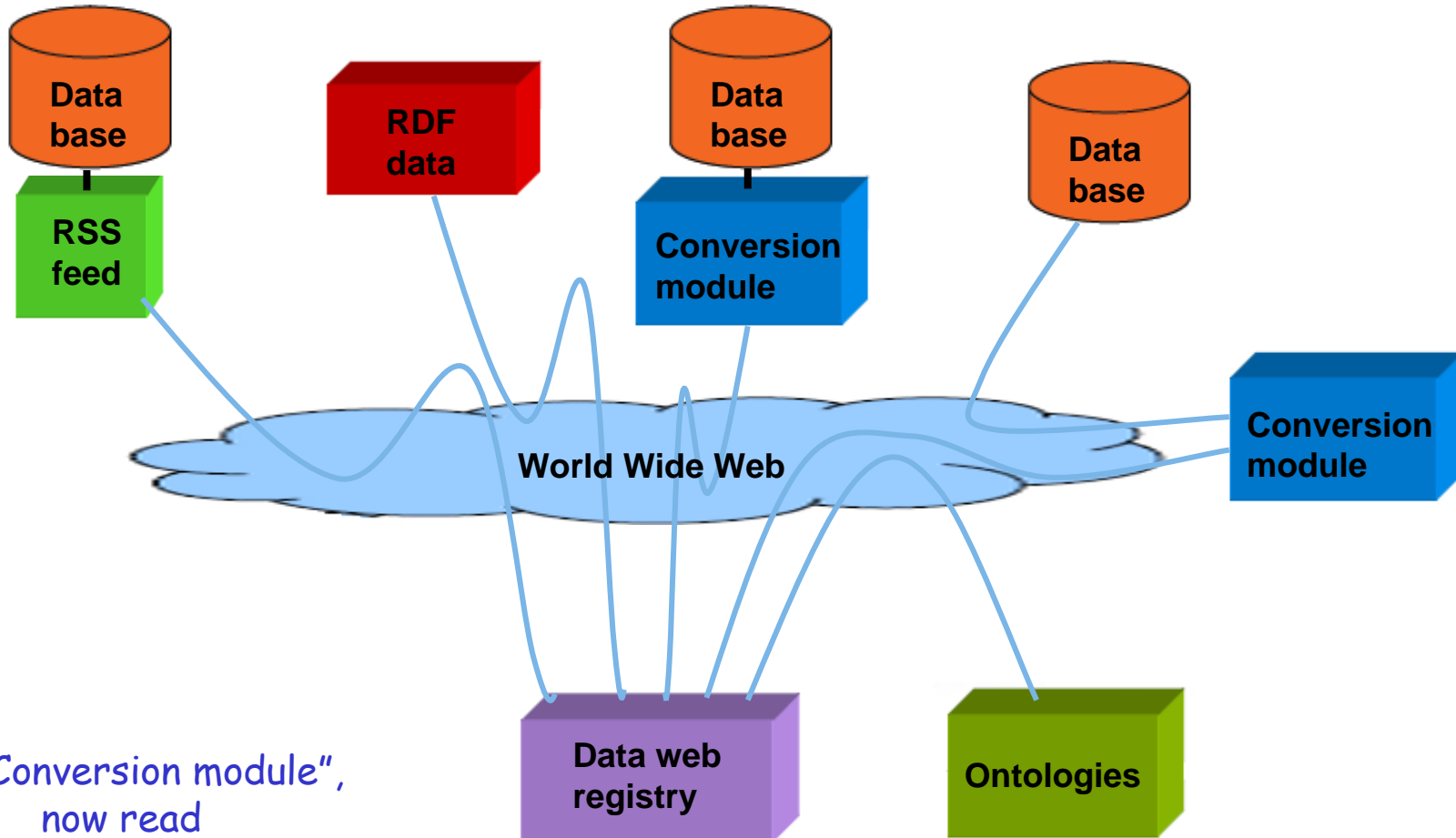
**Data Webs:
Web 2.0 Alternatives to Databases**

David Shotton

Database integration - the status quo

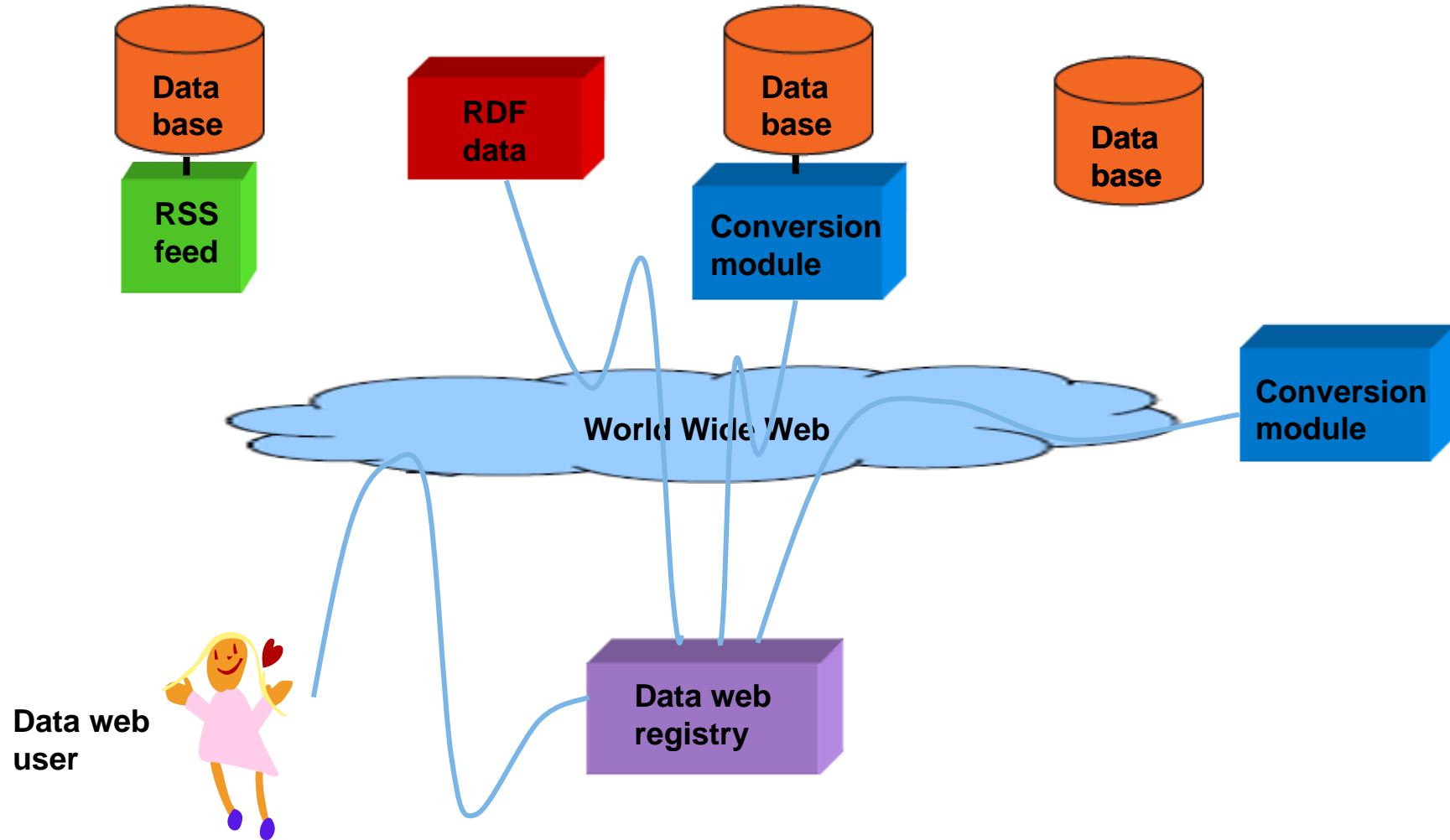


The Data Web Model - data acquisition and indexing

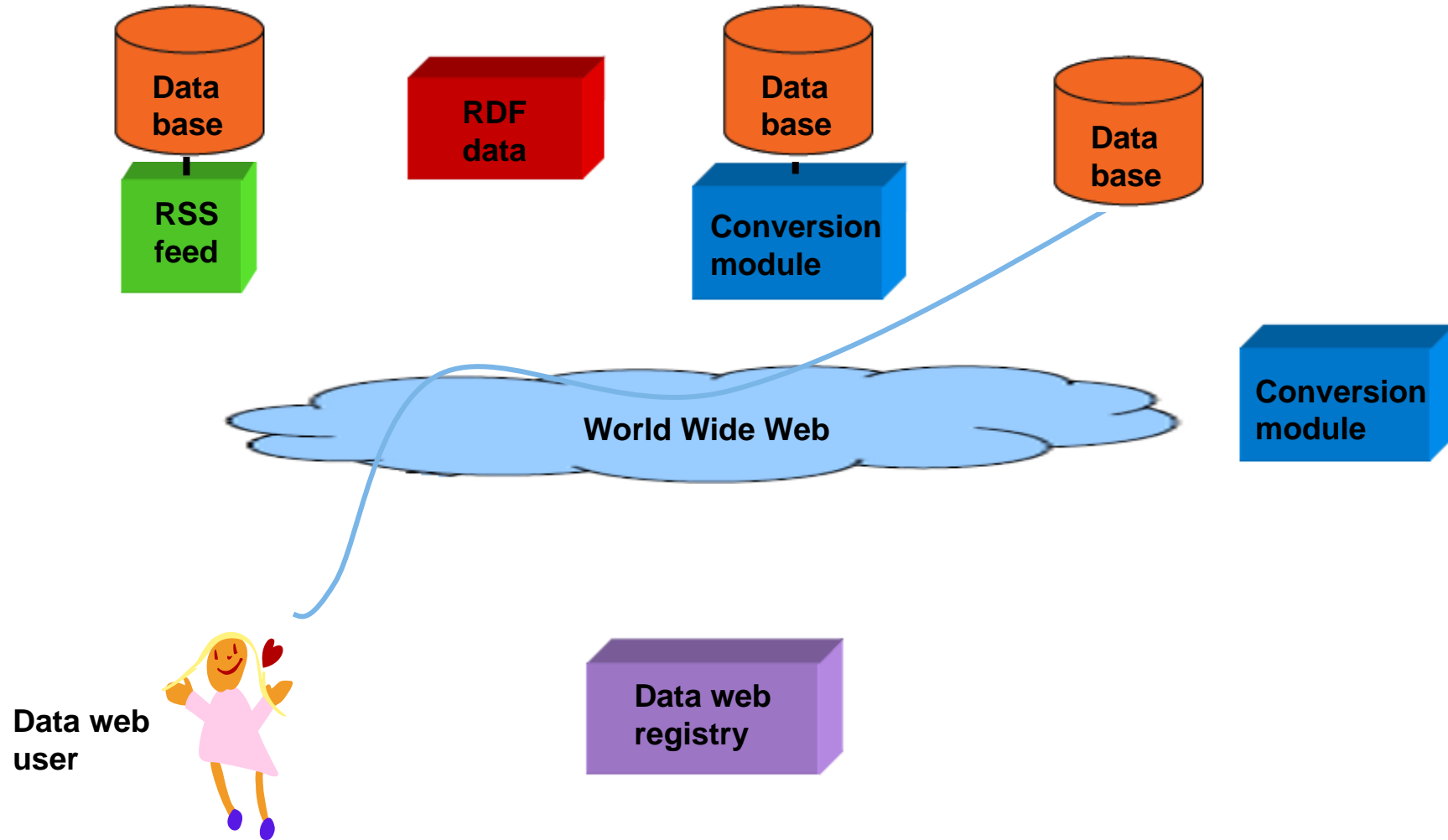


For "Conversion module",
now read
"SPARQL endpoint".
SPARQL was just being
invented in early 2006!

The Data Web Model - user query



The Data Web Model - user referral

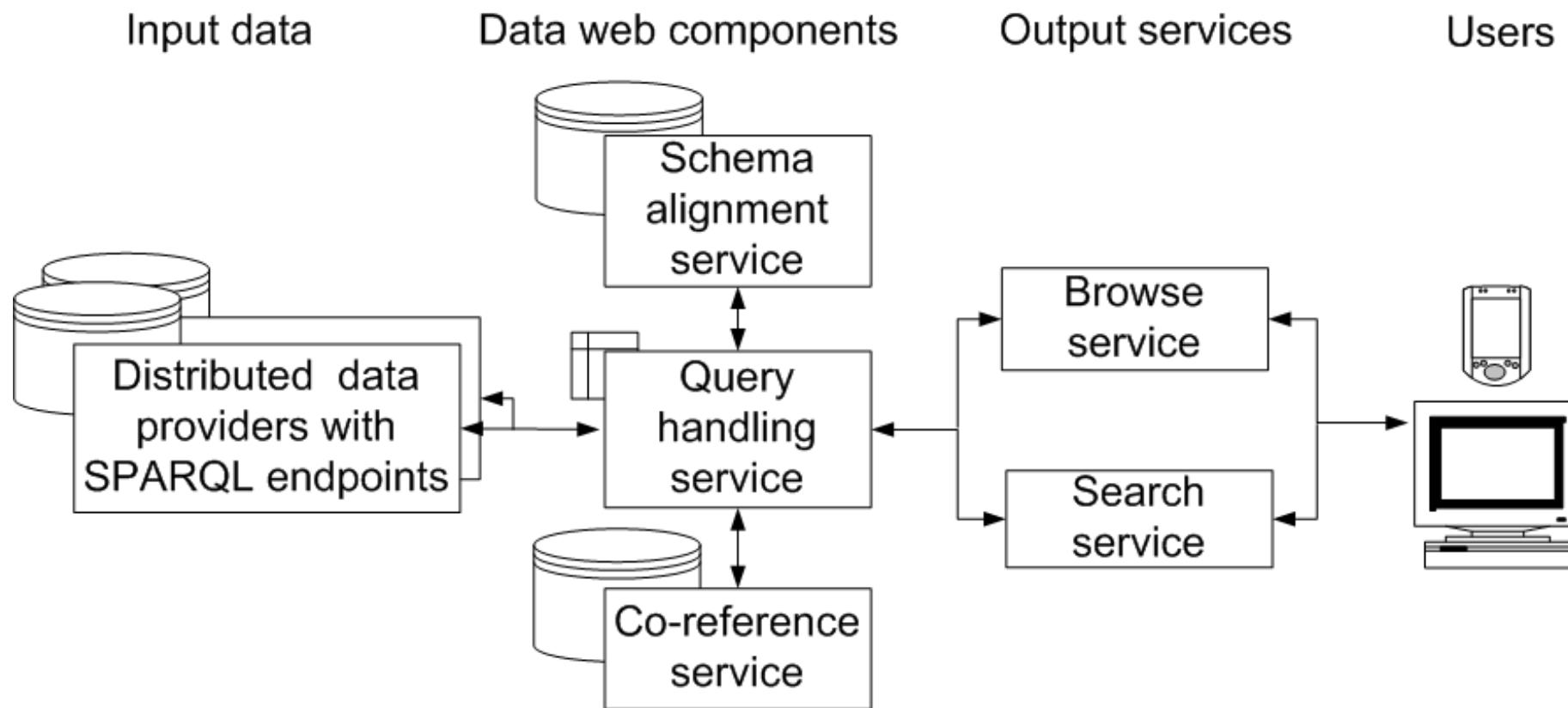


So that was the data web vision in 2006

The challenges of data integration

- **Syntactic differences** between data sources
 - Data are stored in incompatible formats within different DBMSs
 - Solved by converting all data to RDF
- **Semantic differences** between data sources
 - One person's "author" is another person's "creator"
 - Solved by mapping to a common data schema or ontology
- **The co-reference problem**
 - The same entity - for example a particular gene - is known by different names in different databases
 - Solved by creating a co-reference service to disambiguate synonyms

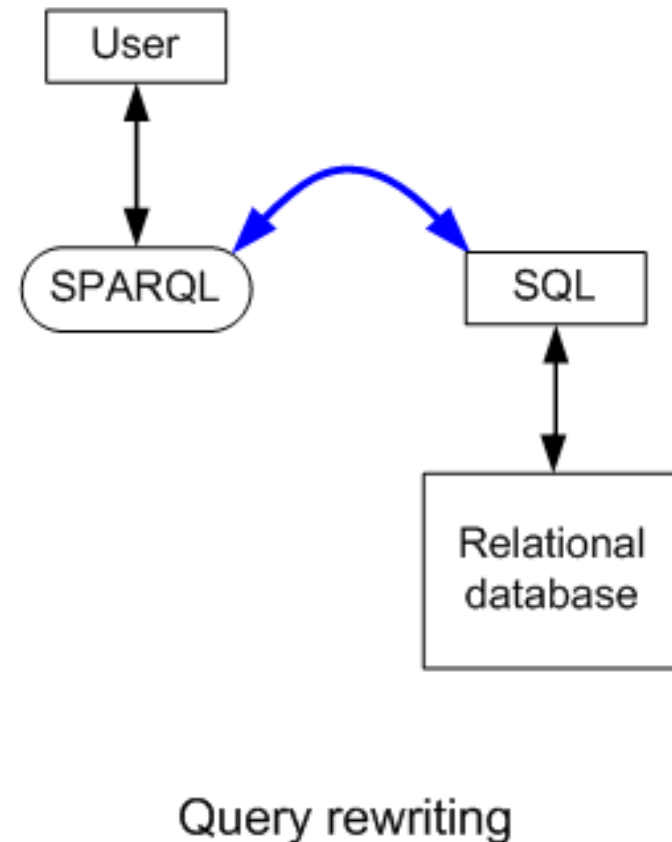
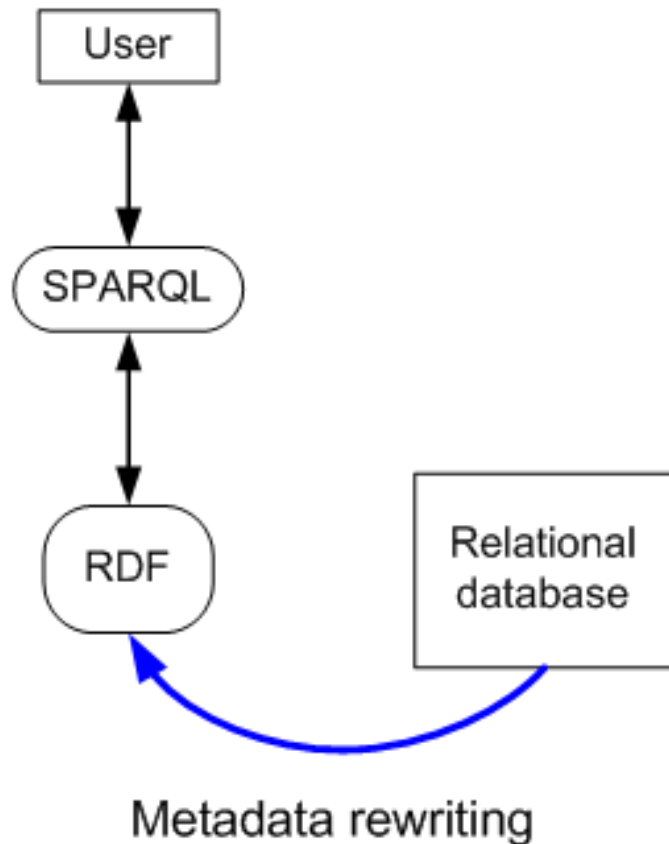
The fundamental components of a data web



Epistemic Networks and GRID + Web 2.0 for Arts and Humanities

Imperial College Internet Centre Wednesday 30 January 2008

Two methods of creating a SPARQL endpoint



- Creation of a local RDF triplestore that **cached** selected source metadata, which are SPARQLed - "**RDF caching**"

- Use of **D2R Server** to rewrite the SPARQL query into the database native query language (SQL) - "**SPARQL virtualization**"

OpenFlyData

An exemplar data webs,
integrating heterogeneous gene
expression data from distributed
bioinformatics sources on the fly

<http://openflydata.org/>

OpenFlyData sources: *Drosophila* gene expression data

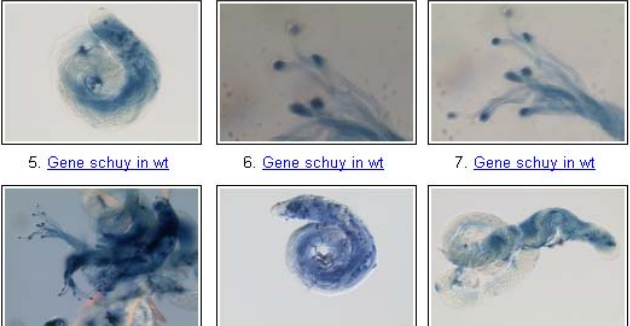
FlyTED: the *Drosophila* Testis Gene Expression Database

Home Browse by Gene Name Browse by Strain Browse by Expression Location Links



Welcome to FlyTED

schuy


- * Browse by Gene Name
- * Browse by Strain
- * Browse by Expression Location
- * Related Links
- * About the Dataset



5. Gene schuy in wt 6. Gene schuy in wt 7. Gene schuy in wt

BDGP In situ homepage - Mozilla Firefox



Patterns of gene expression in *Drosophila* embryogenesis

Release 2 (March 2007)

FAQ User feedback People

Project overview


We use high-throughput 76-well plate RNA *in situ* protocols to determine patterns of gene expression during embryogenesis for *Drosophila* genes represented in non-redundant sets of *Drosophila* ESTs DGC1 and DGC2. At the end of production pipeline gene expression patterns are documented by taking a large number of digital images of individual embryos. The quality and identity of the captured image data are verified by independently derived microarray time-course analysis of gene expression using Affymetrix GeneChip technology (download array data from [here](#)). Gene expression patterns are annotated with controlled vocabulary for developmental anatomy of *Drosophila* embryogenesis. Image, microarray and annotation data are stored in a modified version of Gene Ontology database and the entire dataset is available on the web in browsable and searchable form (see below) or mysql dump can be downloaded from [here](#). So far we examined expression of 6138 genes and documented 3418 of them with 74612 digital photographs. [Summary page](#) summarizes all genes that have *in situ* images.

Search tools

Gene: Enter a gene name to search for *in situ* patterns, use "*" as wild card

Search form for *in situ* patterns by gene name, body part, function, cytology, protein domains

Browse controlled vocabulary used to annotate *in situ* expression patterns



String to look for? schuy e.g. vha, cell adhesion, receptor, aquaporin, adenylate, CG1147, pnt

Searching for SCHUY through 18783 annotations produced 1 hits



[schumacher-levy \(schuy\) FBgn0036925; FBgn0036925; 1626032_at](#)

Accessions: NP_649166.1; NM_140909.

BDGP gene expression has embryonic *in situ* data for CG17736: 11 pix of staining in 2 body parts.

VDRG has 2 RNAsi stocks (nable,lethal) for FBgn0036922 available for purchase.

Tissue	mRNA Signal	Present Call	Enrichment	Affy Call
Brain	1 ± 0	0 of 4	0.00	Down
Head	0 ± 0	0 of 4	0.00	Down
Thoracoabdominal ganglion	1 ± 0	0 of 4	0.00	Down
Salivary gland	3 ± 1	0 of 4	0.04	Down
Crop	0 ± 0	0 of 4	0.00	Down
Midgut	0 ± 0	0 of 4	0.00	Down
Tubule	1 ± 0	0 of 4	0.00	Down
Hindgut	0 ± 0	0 of 4	0.00	Down
Ovary	0 ± 0	0 of 4	0.00	Down
Testis	1140 ± 36	4 of 4	13.90	Up
Male accessory glands	0 ± 0	0 of 4	0.00	Down
Virgin spermatheca	0 ± 0	0 of 4	0.01	Down
Mated spermatheca	1 ± 0	0 of 4	0.01	Down
Adult carcass	1 ± 0	0 of 4	0.00	Down
Larval Salivary gland	0 ± 0	0 of 4	0.01	Down
Larval midgut	1 ± 0	0 of 4	0.01	Down
Larval tubule	0 ± 0	0 of 4	0.00	Down
Larval hindgut	0 ± 0	0 of 4	0.00	Down
Larval fat body	68 ± 43	3 of 4	0.80	None
Whole fly	81 ± 7	4 of 4		

FlyAtlas

FlyBase Gene Report: Dmel:schuy - Mozilla Firefox

FB0009_02, released February 20, 2009

FlyBase Gene Dmel:schuy

Home Tools Files Species Documents Resources News Help Archives Jump to Gene Go

Profile Manager

General Information	
Symbol	Dmel:schuy
Name	schumacher-levy
Feature type	protein_coding_gene
Gene Model Status	Current
Characterization Status	Uncertain(3)
Genomic location	
Chromosome (arm)	3L
Cytogenetic map	7E08-7E09
Genomic maps	Recombination map
Sequence location	3L:19,979,177..19,981,172 [+]
FlyBase GBrowse	
mRNA/EST/GBrowse	
Decorated FASTA	
Get genome region	
Gene region	
Get FASTA	

Summary Information

Automatically generated summary: The gene *schumacher-levy* is referred to in FlyBase by the symbol *Dmel:schuy* (CG17736, FBgn0036925). It is a protein_coding_gene from *Drosophila melanogaster*. It has the cytological map location 7E08. Its sequence location is 3L:19979177..19981172. Its molecular function is described as zinc ion binding. The biological processes in which it is involved are not known. One allele is reported. No phenotypic data is available. It has one annotated transcript and one annotated polypeptide.

See sections below for more information

External Summaries

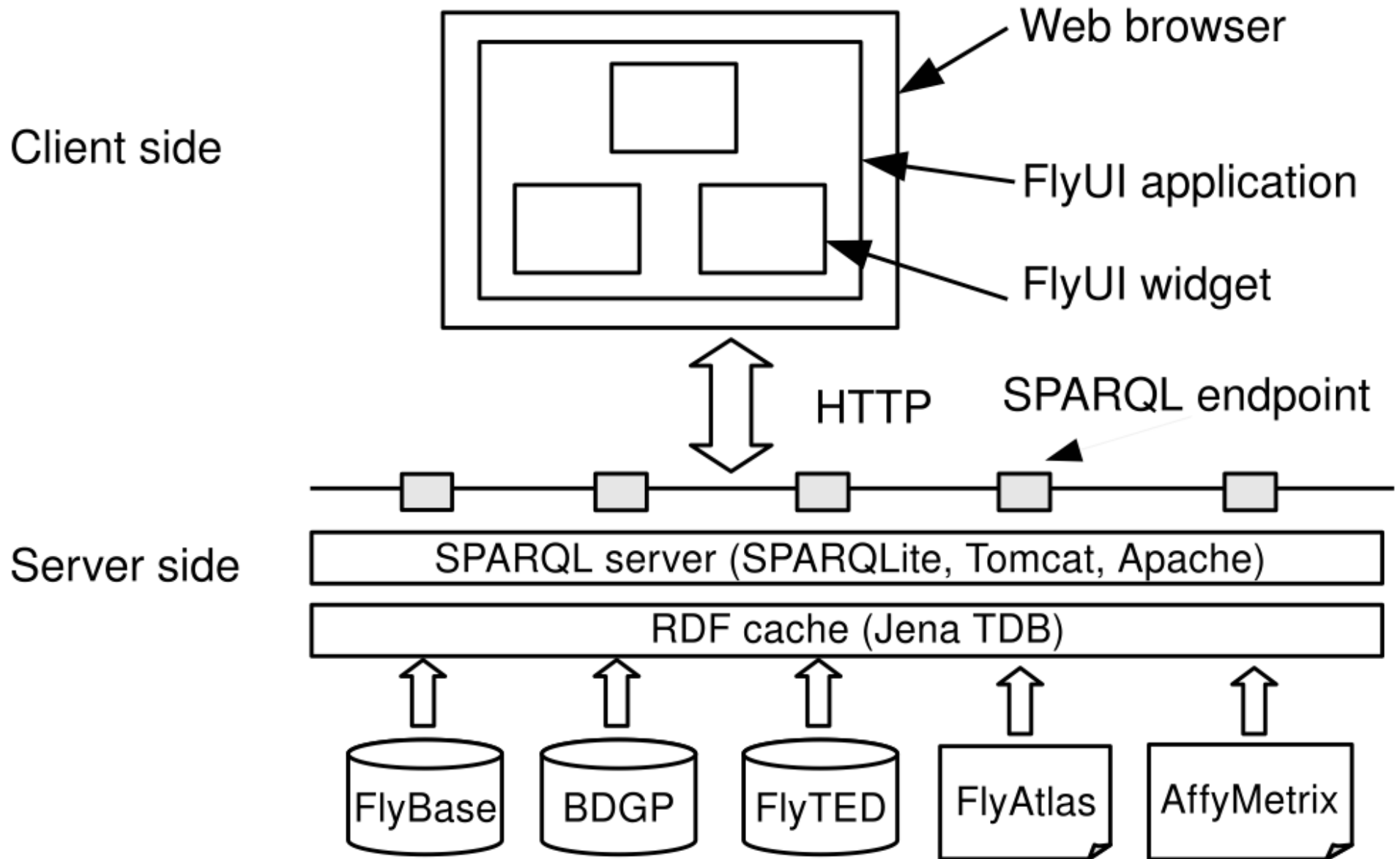
Technical approaches for our creation of data webs

- We use **the Web as the platform**
- We use **W3C Semantic Web tools and standards**:
 - All identifiers are mapped to **URIs**
 - **RDF** is used as the standard format for sharable metadata
 - The RDF query language **SPARQL** is used for data web queries
 - A SPARQL Web service endpoint is made on each data resource
- We employ **open source software components**
 - Ubuntu Linux, Apache web server, Tomcat application server
 - Jena TDB database as the RDF triplestore
 - Yahoo's YUI library for Javascript interfaces
- We use 'agile' user-led development techniques

New functionalities developed to create OpenFlyData

- **Resource-specific ontologies** to support transformation of data from FlyBase, FlyAtlas, BDGP and FlyTED to RDF
 - available at <http://openflydata.googlecode.com>
- **SPARQL endpoints** for each data resource, from which RDF can be obtained in response to SPARQL queries
- **FlyUI** (based on the YUI library, <http://developer.yahoo.com/yui/>) - a library of JavaScript widgets providing re-usable user interface components for displaying *Drosophila* gene expression data
 - available at <http://flyui.googlecode.com>
 - these JavaScript applications run in a Web browser and fetch RDF data asynchronously over HTTP from the SPARQL endpoints
- **SPARQLite**, an implementation of the SPARQL protocol that avoids some performance problems when accessing the underlying triple store
 - available at <http://sparqlite.googlecode.com>
- Use of **FlyBase** (<http://flybase.org/>) to disambiguate gene names

The OpenFlyData architecture diagram



Same data integrated in a single OpenFlyData window



- Query SPARQL endpoints established on cached RDF data

Search *D. melanogaster* Gene Expression Data by Gene

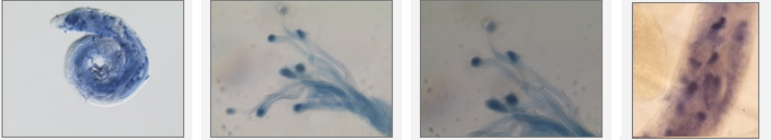
gene name:

E.g. schuy, CG17736 or FBgn0036925 (case doesn't matter)

gene expression levels by tissue					
found 1 matching probe from flyatlas.org (retrieved on 2008-09-16) for gene schuy ...					
probe	tissue	mRNA signal	present call	enrichment	affy call
1626032_at	whole	81.9 ± 7.9	4 of 4	-	-
	brain	1.2 ± 0.6	0 of 4	0	Down
	head	0.4 ± 0	0 of 4	0	Down
	crop	0.8 ± 0.2	0 of 4	0	Down
	midgut	0.6 ± 0	0 of 4	0	Down
	hindgut	0.6 ± 0.1	0 of 4	0	Down
	tubule	1.5 ± 0.7	0 of 4	0	Down
	ovary	0.5 ± 0.2	0 of 4	0	Down
	testis	1140.4 ± 36.8	4 of 4	13.9	Up
	acc	0.9 ± 0.1	0 of 4	0	Down
	l_tubule	0.6 ± 0.2	0 of 4	0	Down
	l_fatbody	68.5 ± 43.9	3 of 4	0.8	None
	ta_ganglion	1.4 ± 0.5	0 of 4	0	Down
	carcass	1.3 ± 0.3	0 of 4	0	Down
sgland	3.5 ± 1.3	0 of 4	0.04	Down	

in situ hybridisation in embryos
found 8 matching images from fruitfly.org (retrieved on 2008-10-30) for gene schuy (BDGP report: CG17736) ...
stage 11-12
no staining;

stage 13-16
gonad;


references (flybase)
found 8 references from flybase.org (FB2009_02) for gene schuy (FlyBase report: FBgn0036925) ...
Barreau et al., 2008, Development 135(11): 1897--1902 Post-meiotic transcription in Drosophila testes. [FBrf0205264]
Dickson et al., 2007.7.18, RNAi construct and insertion data submitted by the Vienna Drosophila RNAi Center RNAi construct and insertion data submitted by the Vienna Drosophila RNAi Center [FBrf0200327]
Benson et al., 2006, A. Dros. Res. Conf. 47: 494A The Drosophila testis gene expression database. [FBrf0189116]
Benson et al., 2005, Europ. Dros. Res. Conf. 19: GG15 The Drosophila testis gene expression database. [FBrf0184732]
Benson, 2004.2.24, Helping FlyBase: ADRC-10142. Helping FlyBase: ADRC-10142. [FBrf0178789]
Benson et al., 2004, A. Dros. Res. Conf. 45: 398B Epigenetic Regulation by the aly-class Meiotic Arrest Genes. [FBrf0173441]

in situ hybridisation in testes
found 11 matching images from www.fly-ted.org (retrieved on 2008-12-03) for gene schuy (Fly-TED reports: schuy)

schuy in wt schuy in wt schuy in wt CG17736/schuy in wt



CLAROS

A data web integrating
heterogeneous classical art data
from distributed sources

<http://www.clarosnet.org/>

The CLAROS data web

- We have used our experience from OpenFlyData to create a data web integrating access to the world's scholarly information on **classical art**
- This required semantic integration of the distributed, heterogeneous and non-interoperable digital resources held by four CLAROS partners
 - the **Beazley Archive** at the University of Oxford
 - the **Forschungsarchiv für antike Plastik** (FAP) in Cologne
 - the **Lexicon Iconographicum Mythologiae Classicae** (LIMC) in Paris
 - the **Lexicon of Greek Classical Names** (LGPN), also in Oxford
- The alpha demonstration service runs at
 - <http://www.beazley.ox.ac.uk/xdb/asp/clarosHome.htm>
 - try the search term "**Heron**"



CLAROS: A data web for classical art

CLAROS

Classical Art Research Online Services

Search

[Home](#)[CLAROS](#)[Pottery](#)[Gems](#)[Sculpture](#)[Iconography](#)[Antiquaria](#)[Dictionary](#)[Tools](#)

Virtual integration of digital assets on classical art

About us

Partner Institutions:

Beazley Archive,
Oxford

German Archaeological
Institute, Berlin

Lexicon
Iconographicum
Mythologiae Classicae
(LIMC) - Paris, Basel

Research Archive for
Ancient Sculpture,
Cologne

Lexicon of Greek
Personal Names



News

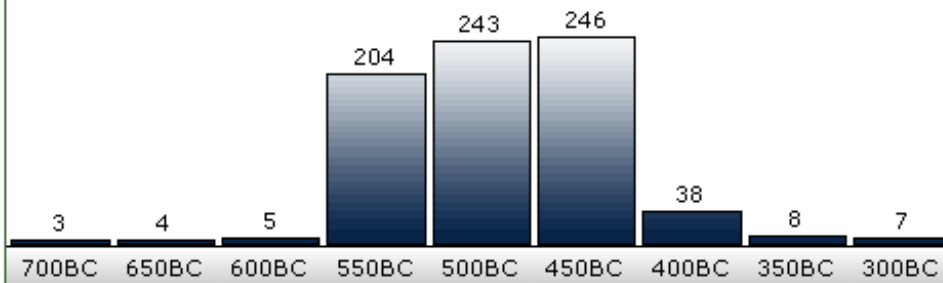
Beazley Archive / LGPN
Pilot Project

Berlin, FIEC Conference, 29
August: proof of concept
launch

Oxford Alumni Weekend,
25 September

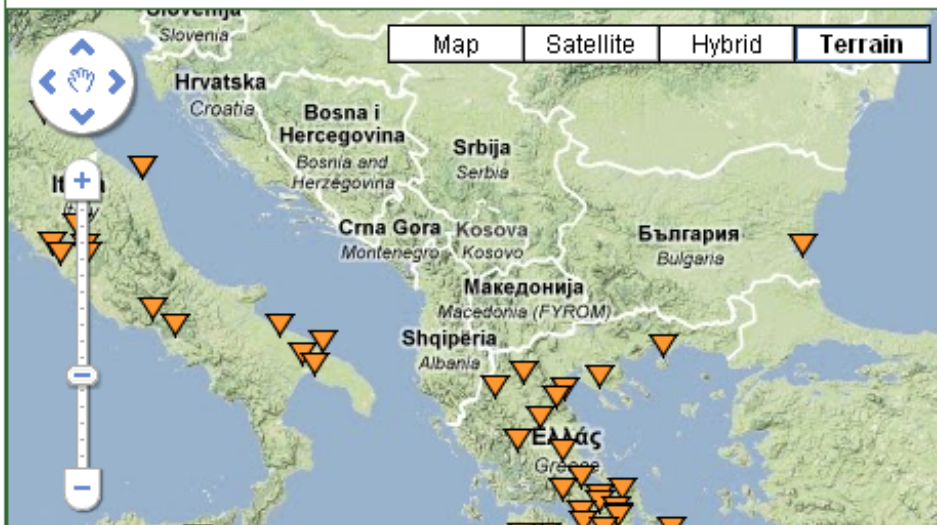
The CLAROS Explorer faceted browse interface

The timeline shows the number of occurrences in each period. Click on the bar to show the distribution within the period or [click here to view the distribution for all periods.](#)



Distribution of heron for all time periods

Click on the marker ▼ on the map and a balloon will pop up with the site name and number of occurrences of the name found at that site.



40008233, Edinburgh Tassie, 2203, EROS HOLDING A HERON BY THE NECK AND RESTING HIS BOW ON THE GROUND, Unpublished Tassie, TRAY 35.2



40000953, Poniatowski, T915, DIOMEDE AND ULYSSES PROCEEDING AS SPIES AT NIGHT TO THE TROJAN CAMP [CONDUCTED BY THE HERON WHICH MINERVA HAD SENT TO GUIDE THEM], KROMOU, Kromos, Prendeville, J.: Explanatory catalogue of the proof-impressions of the antique gems possessed by the late Prince Poniatowski and now in the possession of John Tyrrell, Esq. (1841), 915, Cornelian

Web page search results 1 to 4 of about 4 for heron

[Dexamenos - Classical gems - Gems](#)

... 21mm. GGFR no. 467. Blue chalcedony scaraboid, from Kerch (Crimea). A flying **heron**. ...

A **heron** preening and catching a locust with his foot. Signed Dexamenos. St. ...

<http://www.clarosnet.org/gems/styles/classical/dexamenos.htm> - 8k - 2008-12-11

Other Claros references for this period



Pottery 102155 records



Rundplastik 37905 records



Relief 25488 records

Oxford Vase Search - image recognition software

OxfordVases
Search

In cooperation with

CLAROS

Classical Art Research Online Services

File:

Browse...

or URL:

Upload & Classify

[Help&Hints](#)

[Shape is stamnos](#) (click for shape timeline).

Uploaded vase-image



Foreground separation



Top 3 matches



('stamnos', '')



('stamnos', '')



('stamnos', '')

Matches in Beazley Archive, containing 122193 images. Searched in 0.04513 seconds.

- Developed by **Andrew Zisserman** and his student **Florian Schroff** in the Department of Engineering Science, based on Beazley pot images

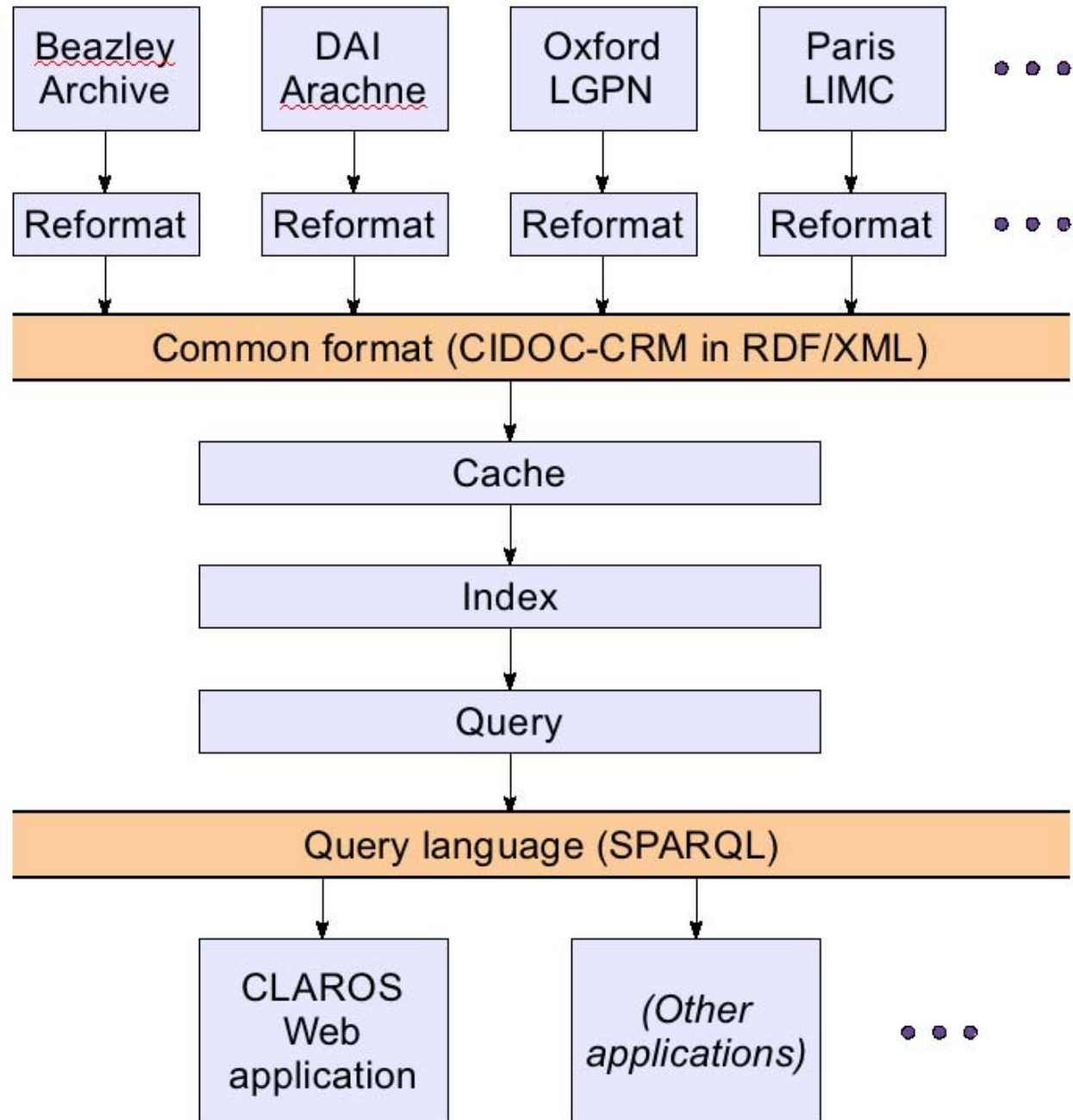
The solutions to the challenges of CLAROS data integration

- **No changes** are required to the databases of the individual sources
- Semantic differences between data sources are resolved by mapping *selected* metadata from each source to **CIDOC-CRM**
- Syntactic differences between data sources are resolved by **converting the selected metadata to RDF**, accessed from a single triple store using **SPARQL**
- The co-reference problem - the same thing being known by different names, e.g. **Heracles, Herakles, Hercules, Ηρακλής** - is resolved (in part) by using the **Lexicon of Greek Personal Names** that records synonyms

CLAROS

architectural diagram

- Data from all sources aligned to a single CIDOC CRM model
- Single CLAROS triple store cache contains about 10 million triples
- User interface is via the CLAROS Explorer
- Other data sources and new user applications can be added ...



CLAROS implementation

- In converting data to RDF, we have used terms from the **OWL** implementation of CIDOC CRM developed by **Erlangen University**
 - The conversion is currently performed by hand-crafted utilities that extract information from each relational database and write it using CIDOC-CRM terms expressed in RDF/XML syntax
- These data are then processed through a set of **inference rules** to create more uniform structures and some additional terms that facilitate search and discovery, and are re-written as RDF into the Jena TDB triple store
- We then run a **Lucene indexer** over the TDB data using LARQ, an extension of the ARQ query component in Jena, to create an additional index to support free-text keyword searches
 - e.g. finding records whose title or description contains "Herakles"
- The front-end **CLAROS Explorer** is an ASP/.NET application providing a multi-faceted browser view, using SPARQL queries to pull information from the triple store in response to user inputs and selections
- **CLAROS is simply a resource discovery service** using minimal metadata - the user is ultimately directed back to the original data publisher's site for full information about an event, object, place or person of interest

The CIDOC Conceptual Reference Model

- Required reading !

The Dream of a Global Knowledge Network—A New Approach

MARTIN DOERR

Institute of Computer Science, Foundation for Research and Technology Hellas (ICS Forth)
and

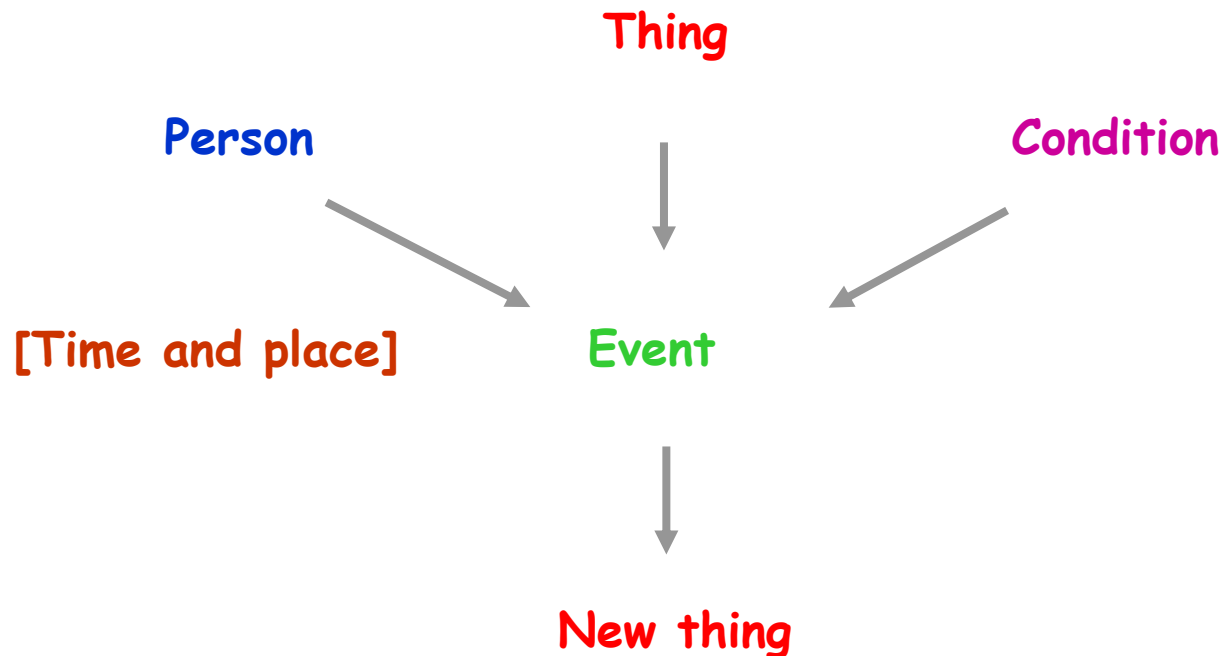
DOLORES IORIZZO

Imperial College, London

ACM J. Comput. Cultur. Heritage 1, 1, Article 5 (June 2008),

<http://doi.acm.org/10.1145/1367080.1367085>

CIDOC-CRM - a simple event-centric data model

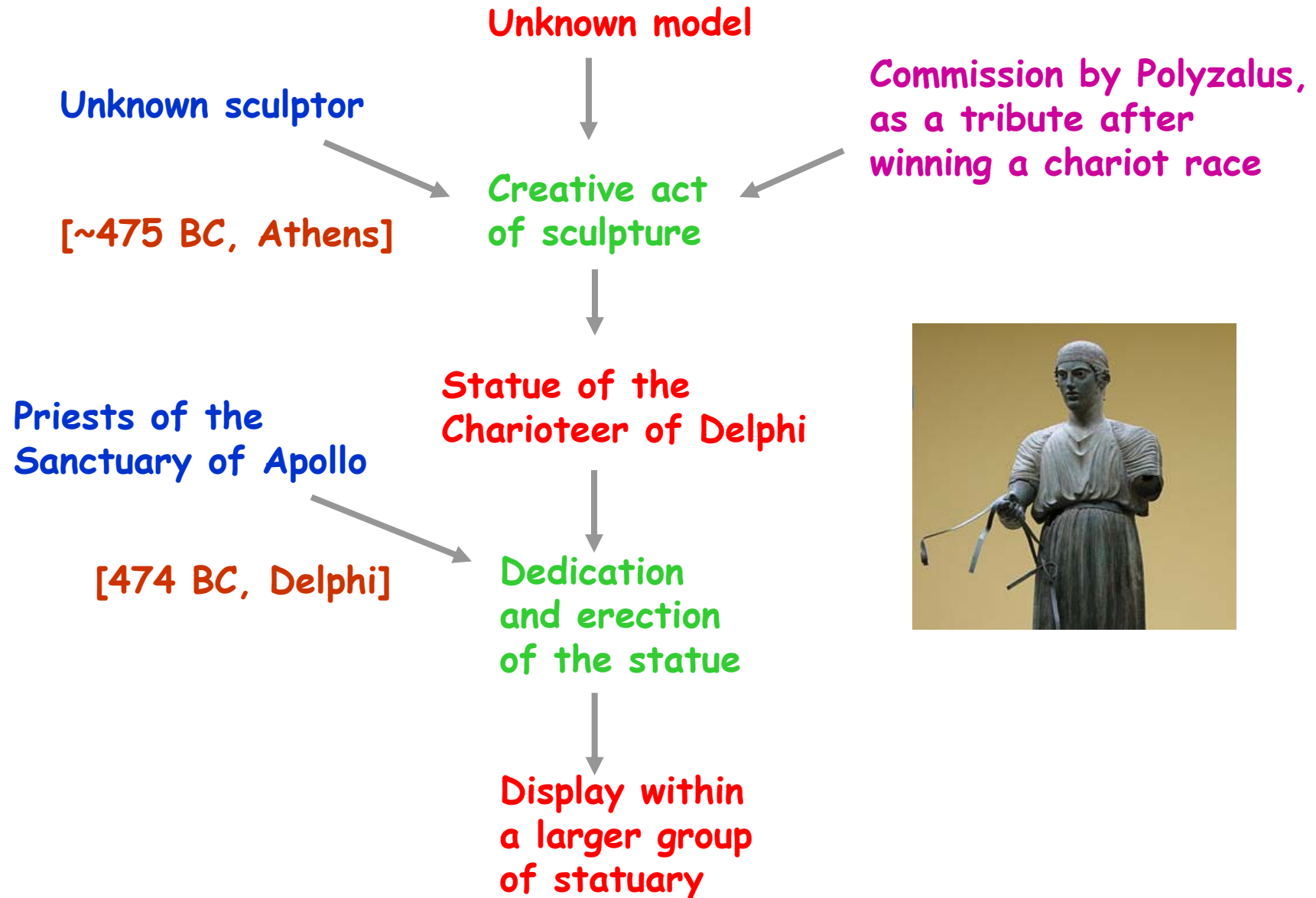


People can be related to objects and conditions through events, that occur at particular times and places

CIDOC-CRM has been developed specifically to describe cultural heritage information

This is exactly the same general data model that we developed for recording experiments and image capture in the BioImage Database !

An event-centric data model for a Greek statue



CLAROS and CIDOC CRM

- We have found CIDOC CRM to be extremely well suited for CLAROS data
- We focused initially on the **CIDOC CRM Core** terms, and employed additional terms as necessary
 - CIDOC CRM Core concisely captures the essential event-mediated structure required to describe the complex provenance of artefacts and their relationships with key events, people, places and times within a single, consistent and surprisingly simple basic framework
- We have found that the diversity of expression possible within CIDOC CRM represents subtle differences of meaning that require a close reading of the CIDOC CRM specification to distinguish
- The necessary complexity of the resulting RDF/XML is mostly invisible to developers, and entirely hidden from users
- We have found the CIDOC CRM "E55.Type" system particularly useful to permit faceted/drill-down queries, e.g. restricting results by the **shape** of a pot
- Once the mappings of relational database terms to CIDOC-CRM had been achieved, running the physical data conversions was trivially easy

The Erlangen OWL implementation of CIDOC CRM

- The OWL implementation of CIDOC CRM by Erlangen University is good, provides a well-maintained URI-based vocabulary for using CIDOC-CRM
- Their on-line documentation is particularly useful
- However, we have concerns that this implementation does not use absolute resolvable URIs
 - The URIs used tend to vary from release to release of CIDOC CRM
 - This is not helpful when trying to use CIDOC CRM in the wider context of the Web of Linked Data
- Additionally, some of their extensions of CIDOC CRM have been provided in a way that makes the extended items opaque to applications that don't understand those extensions

Our CLAROS extensions to CIDOC CRM

- We have made a **very few** extensions to CIDOC CRM to support the requirements of our system, in particular some additional RDF vocabulary for time metadata relating to **imprecise periods and eras**
 - e.g. `claros:not_before` and `claros:not_after`, applied to a `crm:E61.Time_Primitive` object
 - This allows us to capture partial or imprecise quantitative information that is not expressed by a `crm:has_PrimitiveTime` property
- We have undertaken these extensions minimally and in a principled manner, so as **not to obscure** the available CIDOC CRM-encodable information
 - An application that knows only about CIDOC CRM, but does not understand the added CLAROS vocabulary, is still able to access all the CIDOC CRM-encoded information
- Thus we have *not* replaced general properties with extension properties that more precisely captures the intended meanings
- In practice, this means that new properties have been introduced as **OWL datatype properties** on the primitive value objects

Ontologies for Sharing, Ontologies for Use

- In general terms, this approach follows that recommended in our 2005 paper *Catton and Shotton, Ontologies for Sharing, Ontologies for Use*
 - http://imageweb.zoo.ox.ac.uk/pub/2009/publications/Shotton&Catton_2005_Ontologies_for_Sharing,_Ontologies_for_Use.pdf
- The CIDOC CRM standard provides the common 'ontology for sharing'
- CIDOC CRM plus our CLAROS extensions forms our local 'ontology for use'

The CLAROS vision

- The CLAROS system will
 - make cultural assets more widely accessible
 - for scholars
 - for museums and their visitors
 - for students of all ages and abilities
 - for the public at large
 - provide an excellent international model of data federation
 - transform classical art scholarship
- The availability of the **LIMC multi-lingual thesaurus of classical art** will allow CLAROS to be extended to accommodate searches in a variety of languages, to suit users' preferences
- **Personalization** will also be extended to provide varying levels of detail, depending on whether the user is a school child, a student or a scholar



Exhibitions@Home

- The ability to pull together images and information on art objects around the world opens new possibility, for example of creating Exhibitions@Home on any desired subject . . .

- . . . such as depictions of Aphrodite



Aphrodite riding a swan, from Kameiros, Rhodes

- This would permit individuals or schools to experience classical art, with information in their own language and at their own chosen level of complexity, irrespective of museum proximity or geographical locality



Fountain of Aphrodite in Mexico City



Aphrodite, Eros and Pan Delos, 100 BC National Museum Athens

CLAROS challenges and opportunities

Challenges

- We are using SPARQLite to query the Jena TDB triple store
- Nevertheless, some queries, particularly those that require sorting the entire dataset, are 'expensive', resulting in slow performance
- We propose to research solutions to that issue by developing multiple LARQ indexes over the RDF data, enabling the functionality of the triple store more closely to resemble that of a relational database
- Once performance is acceptable, CLAROS Explorer will be made public

Opportunities

- CLAROS is now accepting new partners, bringing complementary data to expand the coverage of classical art objects and their understanding
- Enquiries from potential partners should in the first instance be addressed to Donna Kurtz of the Beazley Archive
<donna.kurtz@beazley.ox.ac.uk>
- In time, CLAROS might become a global catalogue of classical art

Acknowledgements



- My colleague **Graham Klyne**, with whom my initial image web ideas were developed, who has undertaken **all** the development work for CLAROS using CIDOC-CRM
- **Alistair Miles** and **Jun Zhao** who developed OpenFlyData
- **Donna Kurtz**, who had the vision to realize that our data web ideas might be useful for data integration between CLAROS partners
- Her Beazley colleague **Greg Parker** for designing the CLAROS Explorer user interface
- **Florian Schroff** and **Andrew Zisserman** who created the Oxford Vase Search
- **JISC** for funding OpenFlyData and Oxford's **John Fell Fund** for funding CLAROS



Beazley Archive
Classical Art Research Centre



JISC



end