

UK e-Science All Hands Meeting
Nottingham
September 20th 2005

Using the Semantic Web to address problems inherent in biological information management

David Shotton

Oxford e-Science Centre

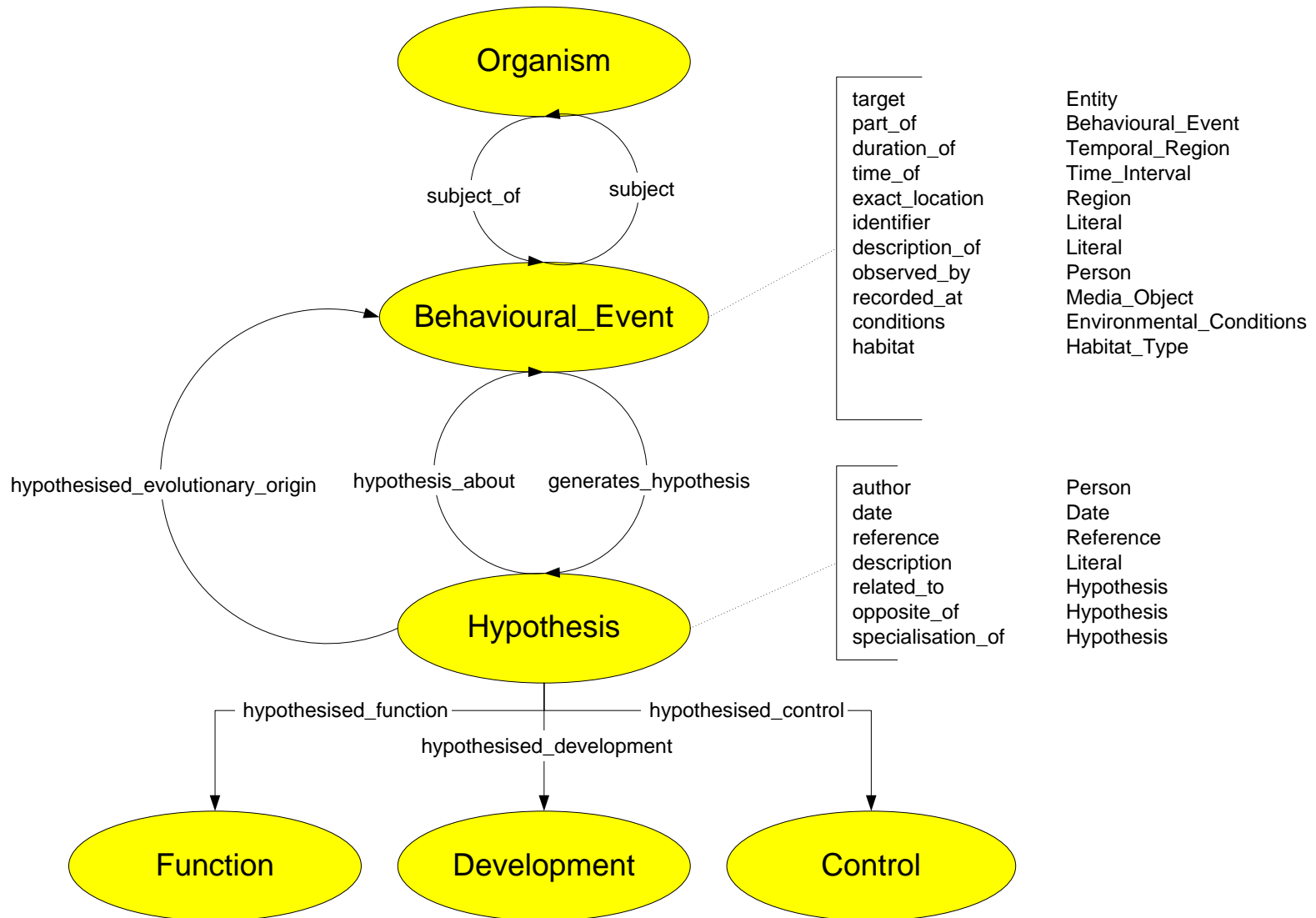


Image Bioinformatics Research Group
Department of Zoology
University of Oxford
Oxford OX1 3PS, UK

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



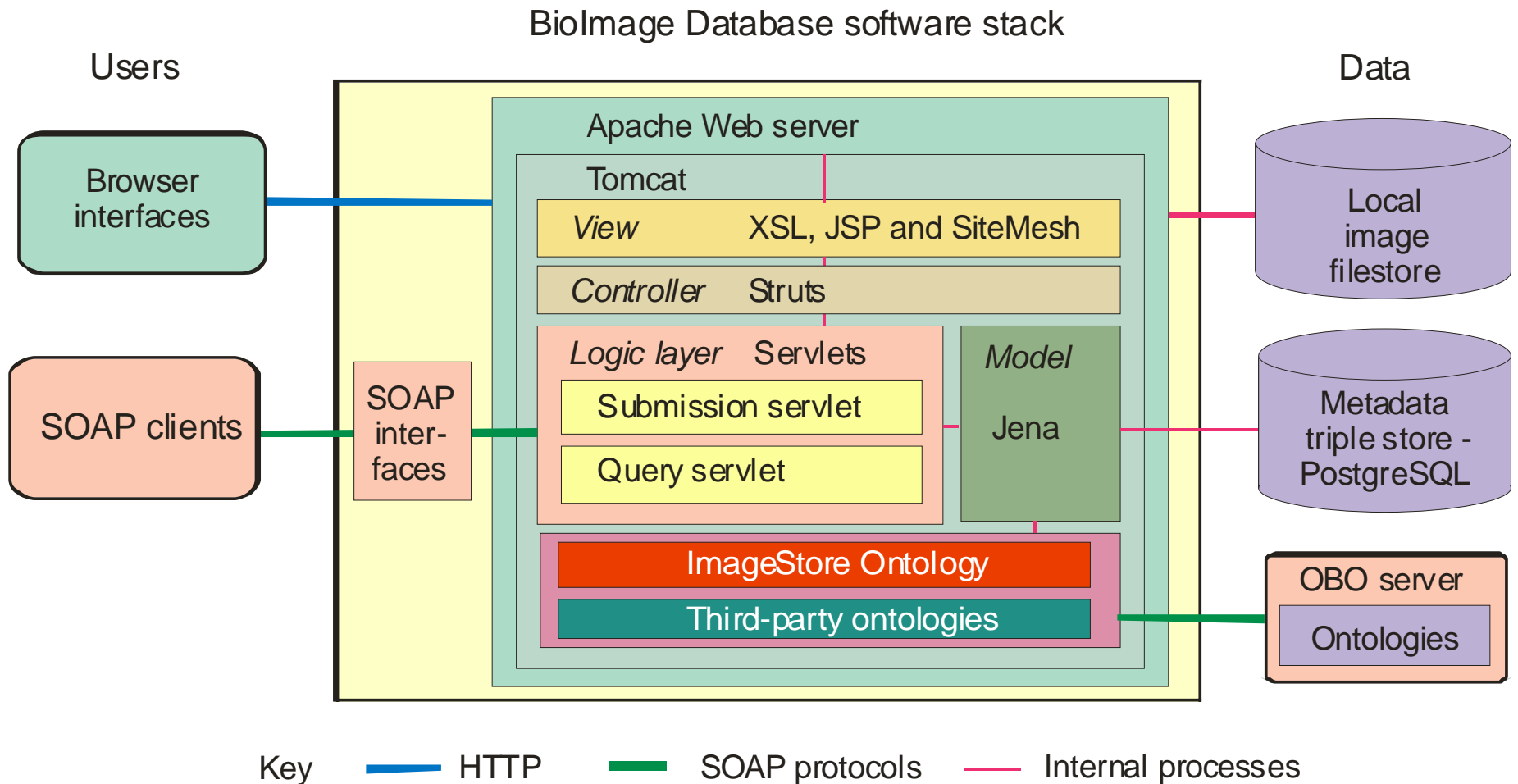
Separating events from functions in animal behaviour



BioImage – a Semantic Web image database



- Open source components, standards compliant and ontology driven




A typical systems biology workflow



- Sperm maturation in normal and 5 meiotic arrest mutants of *Drosophila*
- Affimetrix DNA array screening for altered expression of 15,000 genes
- Affi-data subjected to statistical analysis and clustering
- 10% of genes selected, after consulting literature and FlyBase, and assayed by real-time PCR to check expression and generate primers
- Primers used to create probes for *in situ* hybridization experiments to reveal sites of gene expression in wild type and mutant testes
- Several *in situ* images collected from each specimen
- The images, their annotations and interpretations, together with the 'upstream' PCR and array data, *and* the reasons for choice of genes to study, all need to be recorded in the [Drosophila Testis Gene Expression Database](#) as a public research resource





Repurposing BioImage as a semantic data marshal

- We are now discovering the usefulness of the BioImage Database system as a private laboratory information management and knowledge integration system – a semantic data marshal – quite apart from its use in publishing image data
- We can handle all the additional data types simply by adding the appropriate ontology and by making the data available in RDF
- Using the ontology-driven submission interfaces, metadata can be entered into a Web form – a non-threatening user interface – and saved as RDF. This eases the task of migrating to RDF data
- No major changes are required for this fundamental and unanticipated extension of the usefulness of the BioImage Database
- It all flows from the flexible ontology and RDF data structure adopted
- The semantic data marshal permits integration of both personal and Web-derived data. Can this integration idea be taken a step further?

The biological data publication problem

- Conventional databases for scientific publication are in presently crisis. Do we just keep pumping more resources into such centralised database systems, or is there an alternative that could be explored?
- We have seen and solved similar types of problem before:
 - telephone switchboard operators and typing pools
 - two examples of scaling problems solved by new technologies
- The Semantic Web enables **distributed primary data publication**
 - primary research data need not be *submitted* to a central database
 - rather they can simply be published to the Web in RDF, in conformance with one or more published ontologies – easy!
- What does this bring? All the advantages and disadvantages of the Web itself: Decentralisation and democratization of publication; insecurity of data on individual servers; built-in scalability; etc.
- Metadata can then be harvested into a central metadata registry that provides the ability to search over all the data, returning URIs

An example - Open Microscopy Web

- The Open Microscopy Environment is an open source laboratory image capture, management and analysis system
- It is designed for local private use, but images and their metadata can be exchanged or published in a standard OME-XML format
- Through our knowledge of the OME data model, which closely resembles that of the BioImage Database, we will, in collaboration with distributed OME users, create an Apache module to extract the metadata from their OME-XML files and make them available as RDF
- We will then harvest this RDF metadata centrally, and index it in a special instance of our database – the OMW Metadata Registry
- Data from independent resources will thus be integrated without effort
- Ontology-enabled ‘smart searching’ of this registry will provide users with URIs linking to the original downloadable OME-XML data files
- Value is added by integrating the data and by providing customised user interfaces and search facilities (‘RDF interpreters’)

Advantages of lightweight RDF data publication

- This radically new way of publishing biological data is not appropriate for data representing universal truths, e.g. genome sequences, which need to be held in a single central database
- But many biological data are of a different type – they are particulars rather than universals, forming unbounded data sets
 - e.g. cell biology images, experimental protocols, observational records of mouse behavioural mutants
- The scope of a distributed database is defined by the ontologies it chooses to integrate. RDF gives the possibility of expanding or changing this scope with unprecedented ease
- The primary data are never owned by the database, but are freely available for use by other presently unforeseen applications, including data aggregation and analysis services
- Control of the primary data is left in the hands of the data creators or their local institutional data repositories

Acknowledgements



Chris Catton

BiImage Development Manager and ImageStore Ontology creator

Simon Sparks

BiImage Software Engineer

John Pybus

BiImage Systems Manager

Graham Klyne

DTGED Development Engineer

Liz Mellings

BiImage Database Curator



European Commission for funding the ORIEL Project - IST-2001-32688

