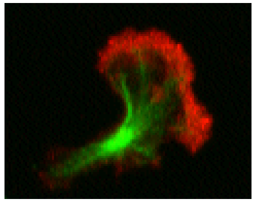


Semantic Interoperability for e-Research in the Sciences, Arts and Humanities

Imperial College
March 30th 2006

Data Webs:

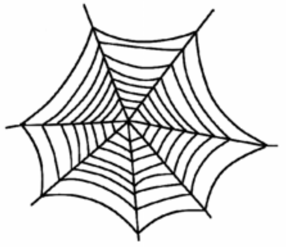
Web 2.0 Alternatives to Databases

David Shotton

Image BioInformatics Research Group
Department of Zoology
University of Oxford, UK

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

bioimageweb



Outline

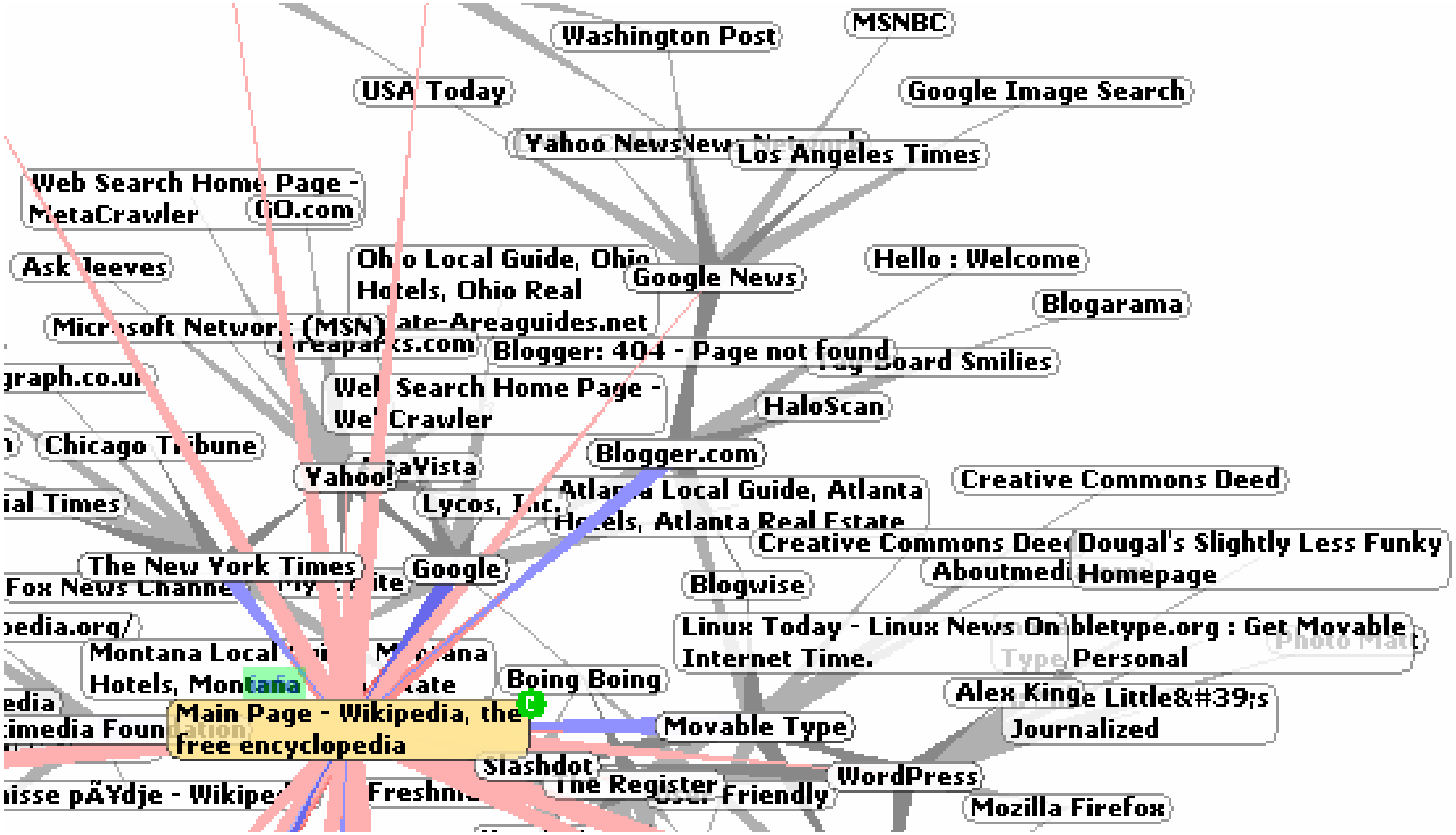
- The Web as we know and love it
- Metadata, ontologies and the semantic web
- Databases . . . and their limitations
- The nature of biological data and biological databases
- Database integration
- The concept of a data web
- The BioImageWeb Project
- Paradigm traps, ‘Web 2.0’, and the social tagging for scientific images

Themes

- The nature and the limitations of the Web
- The advantages and problems of semantically rich metadata and ontologies
- The notion that solutions to biological data management problems may have generic applicability in the arts and humanities, and even beyond academia

The Web as we know and love it

- There are nearly 80 million web sites in the world with registered domains
- The World Wide Web is a scale-free network of hyperlinks
  - Here are just some of the links from the Wikipedia home page



The Web provides documents for humans to read

- The World Wide Web is familiar as an environment in which
  - publication of documents is cheap and easy
  - linking between them is trivial
- The Web is characterized by
  - lack of control
  - freedom and decentralization of publication
  - distributed data
- Its advantages include:
  - a "missing is not broken" Open World philosophy
  - built-in scalability
- Its disadvantages include
  - lack of quality control
  - lack of consistency
- Differences in data presentation formats make collating information from multiple web pages hard for humans and well nigh impossible for machines

The Web, HTML and meaning

- The World Wide Web transmits documents designed to be viewed by people
- It works because of two fundamental technologies:
  - the Hypertext Transport Protocol (http) that permits packets of information to be transferred in such a way as to enable a global hypertext system, and
  - the Hypertext Markup Language (HTML), that defines tags specifying how information is to be displayed in a Web browser window
- For instance, placing the HTML tags

  ```
  <b>  . . .  </b>
  ```

  around a word or phrase instructs browsers to display it in **bold type**
- Together, HTML and http have enabled the development of the Web – a vast network of interlinked documents
- But it is inflexible, since HTML conveys no meaning about the text it marks up

- Metadata, essential both for resource description and resource location, required a richer environment

The role of metadata and ontologies

From *Towards 2020 Science*, Report by Microsoft Research, March 2005 (available at research.microsoft.com/towards2020science)

- "This 'data about data' is not simply for human consumption, it is primarily used by tools that perform data integration . . ."
- "It is not practical to attempt to capture everything a paper contains – present-day ontologies and data models are nowhere near as expressive as human languages – but in principle, we can provide a useful summary of the bibliographic details, authors, institutions, methods and citations, as well as the main scientific entities (molecules, genes, species and so on) with which the paper is concerned."
- "This, in turn, should enable much more specific searching of, and linking to, the paper in question."

How to classify . . .

“On those remote pages it is written that animals are divided into:

a. **those that belong to the Emperor**

b. **embalmed ones**

c. **those that are trained**

d. **suckling pigs**

e. **mermaids**

f. **fabulous ones**

g. **stray dogs**

h. **those that are included in this classification**

i. **those that tremble as if they were mad**

j. **innumerable ones**

k. **those drawn with a very fine camel's hair brush**

l. **others**

m. **those that have just broken a flower vase**

n. **those that resemble flies from a distance**"

From *The Celestial Emporium of Benevolent Knowledge*, Jorge Luis Borges

Structuring metadata

- Free text tagging, as in the previous example
- A controlled vocabulary (a word list with no internal structure)
- A hierarchical taxonomy of ‘parent-offspring’ *is_a* relationships
    - e.g. a crow is a bird, a bird is a vertebrate
- A thesaurus, in which additional relationships between terms may be defined
- An ontology, in which such relationships are, ideally, defined in such a manner as to permit computers to make semantic inferences and undertake logical reasoning over the data
- A helpful definition of an ontology has been given by Tom Gruber as
    - The formal explicit specification of a shared conceptualisation
- The role of an ontology is thus to facilitate the formal sharing and re-use of knowledge through the construction of an explicit domain model

The Semantic Web

- Tim Berners-Lee’s vision of “the Web of integrated data”

W3C® WORLD WIDE WEB consortium

Leading the Web to Its Full Potential...

- The Semantic Web extends the web by providing a data representation that has both syntactic consistency and a semantic framework, enabling both interoperability and computational inferencing
- It involves three technologies, each resting *hierarchically* on the previous one:
  - The eXtenstible Markup Language (XML) that permits one to define the meaning of terms using XML tags, with XML Schema providing *syntactical* structure
  - The Resource Description Framework (RDF) that permits one to make simple logical statements (subject-verb-object, or entity-attribute-value) written in XML, for describing objects and the relationships between them, with RDF Schema providing *semantic* structure
  - The Web Ontology Language OWL, itself expressed as a set of RDF / RDFS statements, to specify the supporting ontologies that provide *semantic definitions* of the RDF terms

How to make ontological statements using RDF

- An RDF triple might state that a mouse *is_a* mammal, informing the computer that an entity ‘mouse’ is included in the more general category of ‘mammal’

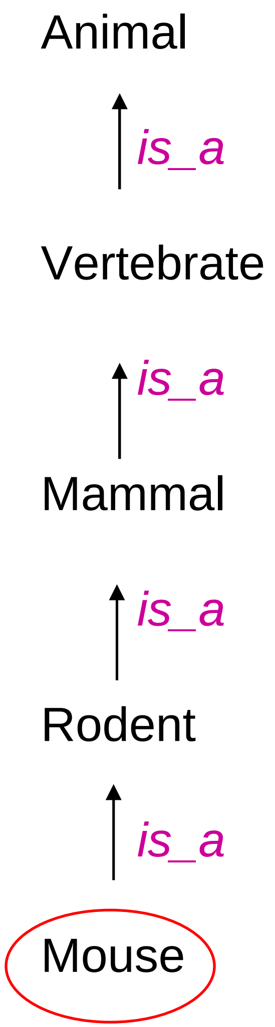
- By using several RDF entity-attribute-value triples referring to the same entity, multiple attributes can be defined:

Subject (Entity)	=	Mouse (class)	or	This mouse (instance)		
Property (Attribute)	=	*is_a*	/	*has_location*	/	*has_identifier*
Object (Value)	=	Mammal	/	*Oxford*	/	*667*

- In RDF, the statement “This mouse is located in Oxford” is simply:

```
<rdf:RDF>
        <rdf:Description rdf:about=“Mouse”>
                    <Location>Oxford</Location>
        </rdf:Description>
</rdf:RDF>
```

An ontology is richer than a taxonomic hierarchy



Animal

is_a

Vertebrate

is_a

Mammal

is_a

Rodent

is_a

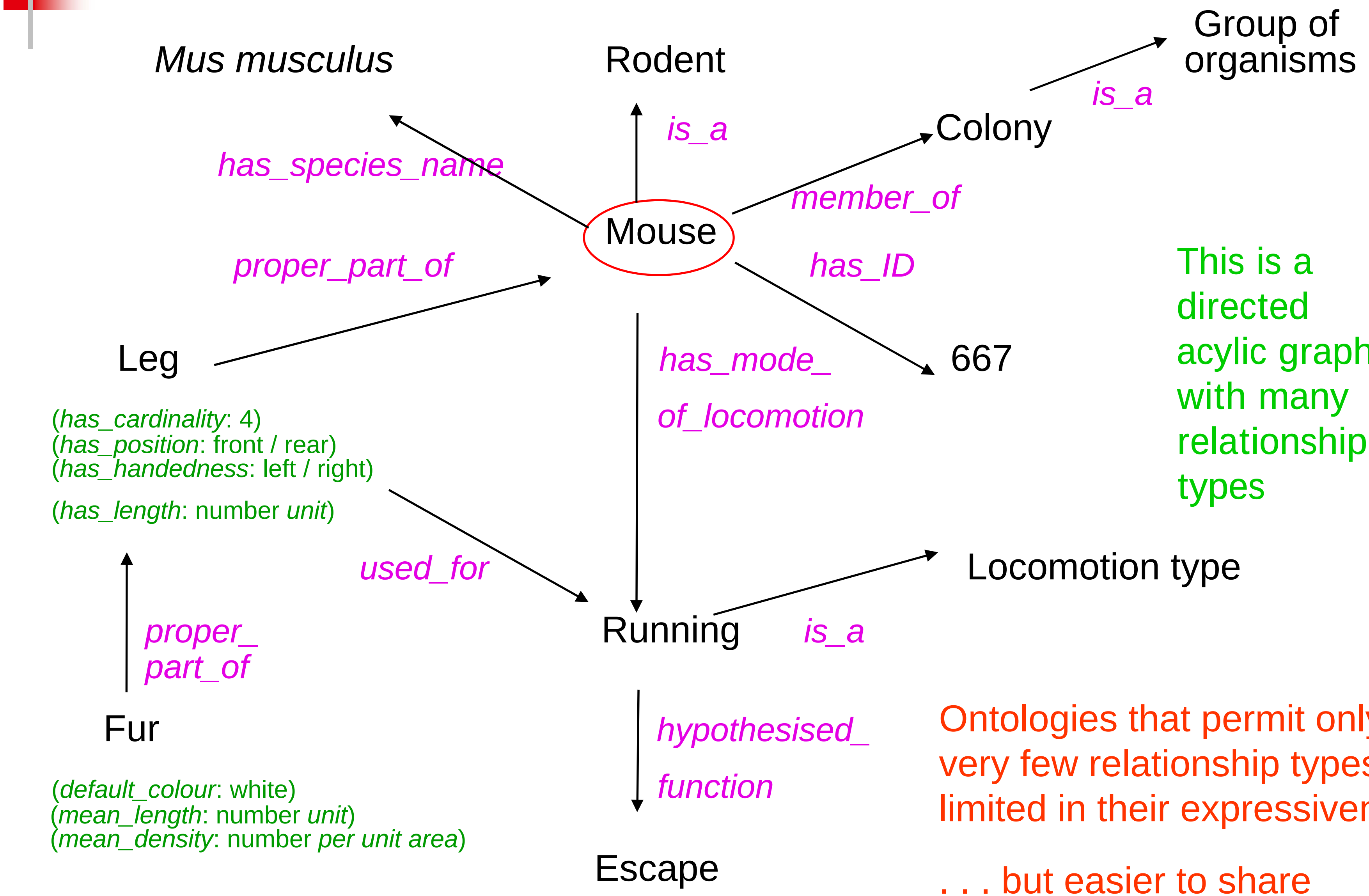
Mouse

Here all the relationships are of a single type, that of being a sub-class, where each sub-class has only one 'parent'. Phylogenetic trees are typical constructs using this relationship

Hierarchies have the advantage that each sub-class (e.g. rodent) inherits all class properties previously defined for its parent class (e.g. mammal), such as the possession of four legs and fur – subsumption

However, in an ontology one can express more complex relationships about a mouse, other than just its taxonomy

A partial ontology of ‘mouse’



Mus musculus

Rodent

Group of organisms

is_a

is_a

Colony

has_species_name

member_of

Mouse

proper_part_of

has_ID

Leg

has_mode_ of_locomotion

667

(*has_cardinality*: 4)
(*has_position*: front / rear)
(*has_handedness*: left / right)

(*has_length*: number *unit*)

used_for

Locomotion type

proper_ part_of

Running

is_a

Fur

hypothesised_ function

(*default_colour*: white)
(*mean_length*: number *unit*)
(*mean_density*: number *per unit area*)

Escape

This is a directed acylic graph with many relationship types

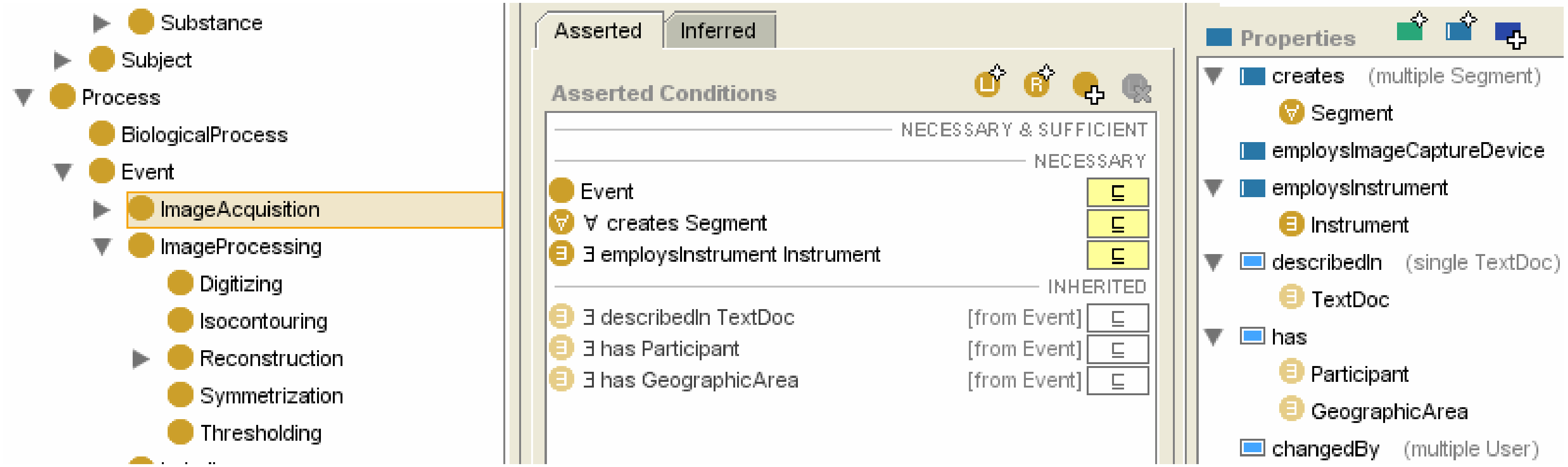
Ontologies that permit only very few relationship types are limited in their expressiveness

. . . but easier to share

How to build an ontology

- Relationships in an ontology take the form of a directed acyclic graph (DAG), in which an entry can have more than one ‘parent’
- An OWL ontology can conveniently be written in RDF, the subject-verb-predicate of an RDF triple equating to an single node-link-node in the DAG
- Tools such as Protégé-OWL make the task of ontology building much easier:

ImageStore



- Substance
- Subject
- Process
  - BiologicalProcess
  - Event
    - ImageAcquisition
    - ImageProcessing
      - Digitizing
      - Isocontouring
      - Reconstruction
      - Symmetrization
      - Thresholding

Asserted | Inferred

Asserted Conditions

NECESSARY & SUFFICIENT

NECESSARY

Event ⊑

∀ creates Segment ⊑

∃ employsInstrument Instrument ⊑

INHERITED

∃ describedIn TextDoc [from Event] ⊑

∃ has Participant [from Event] ⊑

∃ has GeographicArea [from Event] ⊑

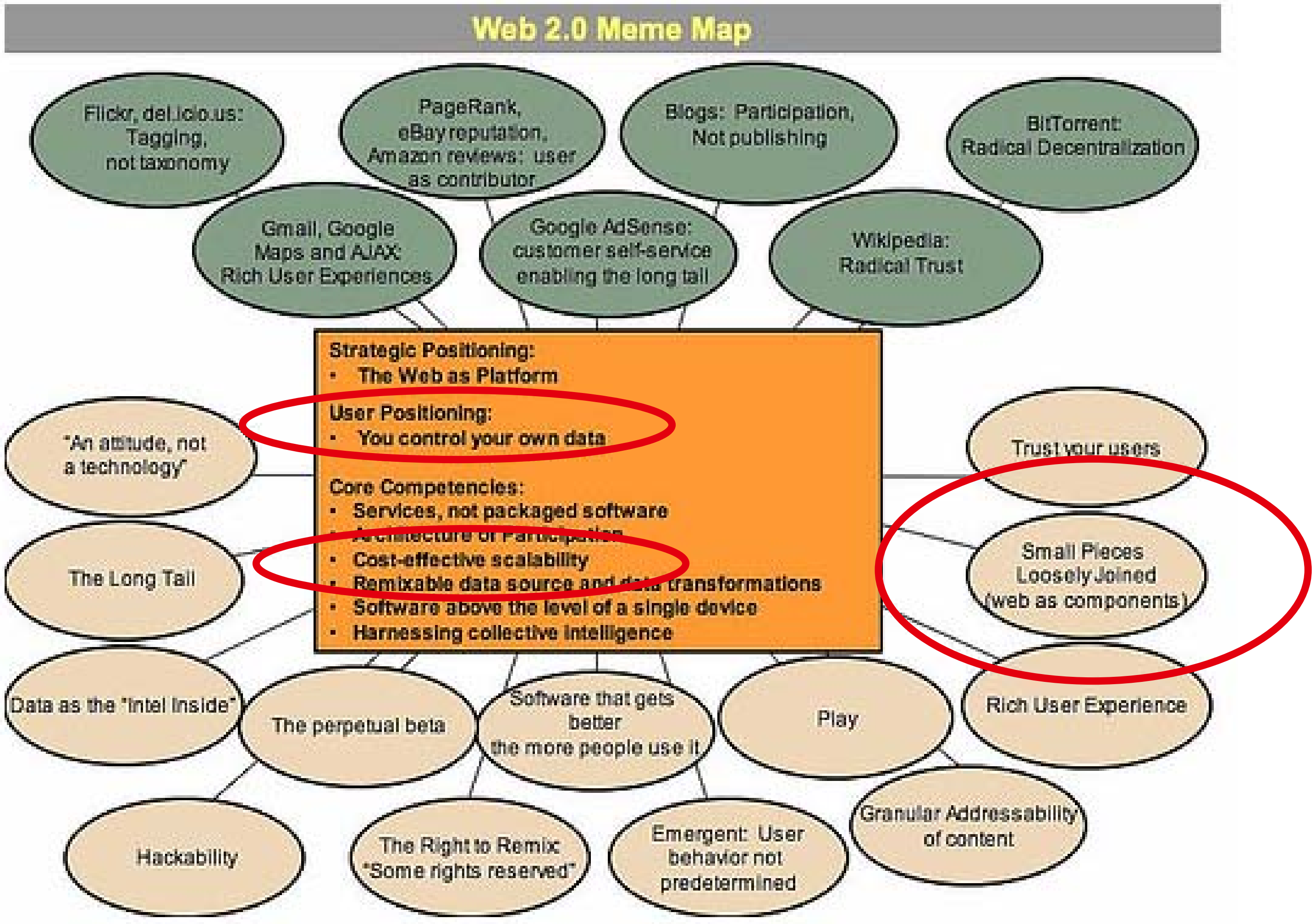
Properties

- creates (multiple Segment)
  - Segment
- employsImageCaptureDevice
- employsInstrument
  - Instrument
- describedIn (single TextDoc)
  - TextDoc
- has
  - Participant
  - GeographicArea
- changedBy (multiple User)

Part of the ImageStore Ontology of the BioImage Database, visualized in Protégé-OWL

Web 2.0 – more a way of thinking

- From Tim O'Reilly's paper "What is Web 2.0" at http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html


Web 2.0 Meme Map

Strategic Positioning:
- The Web as Platform

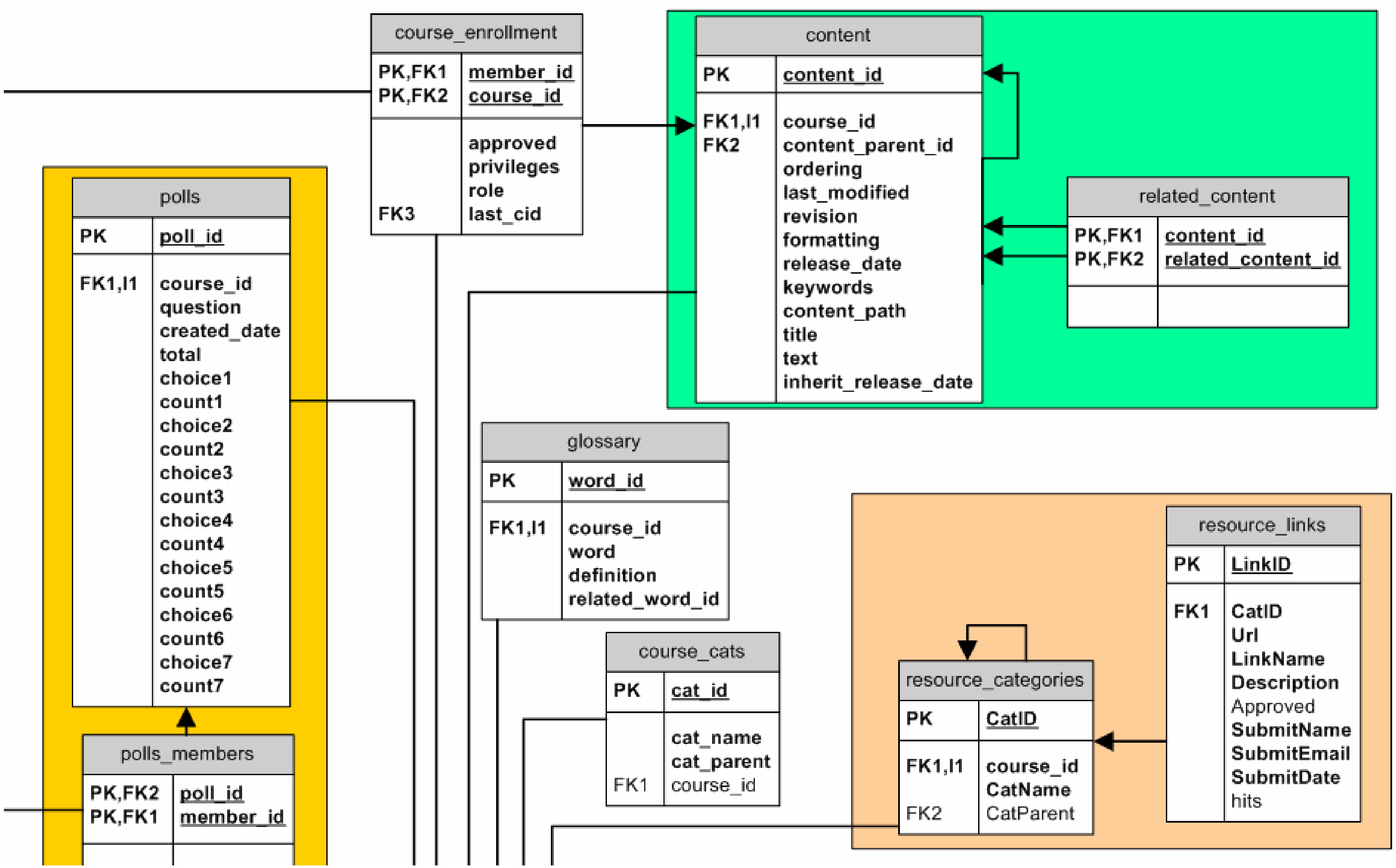
User Positioning:
- You control your own data

Core Competencies:
- Services, not packaged software
- Architecture of Participation
- Cost-effective scalability
- Remixable data source and data transformations
- Software above the level of a single device
- Harnessing collective intelligence

Flickr, del.icio.us: Tagging, not taxonomy
PageRank, eBay reputation, Amazon reviews: user as contributor
Blogs: Participation, Not publishing
BitTorrent: Radical Decentralization
Gmail, Google Maps and AJAX: Rich User Experiences
Google AdSense: customer self-service enabling the long tail
Wikipedia: Radical Trust
"An attitude, not a technology"
Trust your users
The Long Tail
Small Pieces Loosely Joined (web as components)
Data as the "Intel Inside"
The perpetual beta
Software that gets better the more people use it
Play
Rich User Experience
Hackability
The Right to Remix: "Some rights reserved"
Emergent: User behavior not predetermined
Granular Addressability of content

Databases as we know and love them

- Much of human knowledge is stored in relational databases



course_enrollment	
PK,FK1	member_id
PK,FK2	course_id
	approved
	privileges
	role
FK3	last_cid

content	
PK	content_id
FK1,I1	course_id
FK2	content_parent_id
	ordering
	last_modified
	revision
	formatting
	release_date
	keywords
	content_path
	title
	text
	inherit_release_date

related_content	
PK,FK1	content_id
PK,FK2	related_content_id

polls	
PK	poll_id
FK1,I1	course_id
	question
	created_date
	total
	choice1
	count1
	choice2
	count2
	choice3
	count3
	choice4
	count4
	choice5
	count5
	choice6
	count6
	choice7
	count7

polls_members	
PK,FK2	poll_id
PK,FK1	member_id

glossary	
PK	word_id
FK1,I1	course_id
	word
	definition
	related_word_id

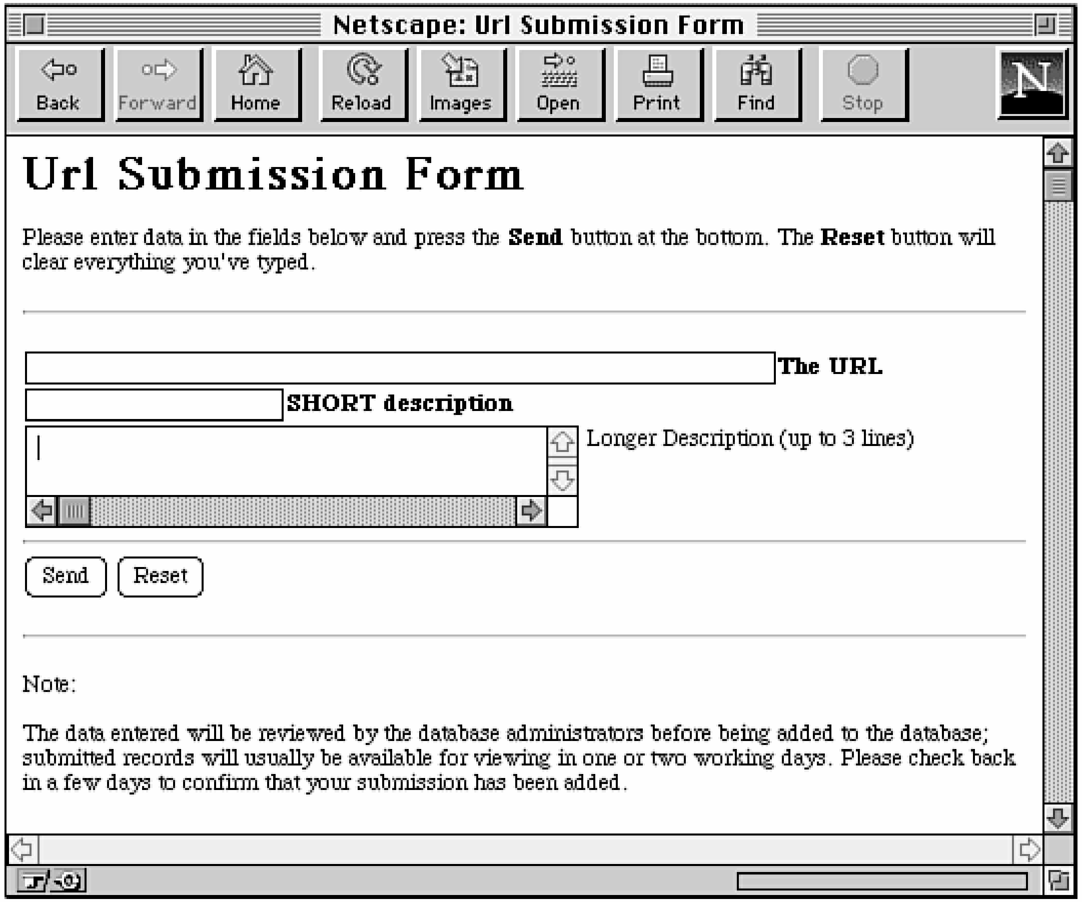
course_cats	
PK	cat_id
	cat_name
	cat_parent
FK1	course_id

resource_categories	
PK	CatID
FK1,I1	course_id
	CatName
FK2	CatParent

resource_links	
PK	LinkID
FK1	CatID
	Url
	LinkName
	Description
	Approved
	SubmitName
	SubmitEmail
	SubmitDate
	hits

Database submission

- Databases are populated by discrete acts of data submission, usually subjected to scrutiny by a database curator to ensure quality



Url Submission Form

Please enter data in the fields below and press the **Send** button at the bottom. The **Reset** button will clear everything you've typed.

The URL

SHORT description

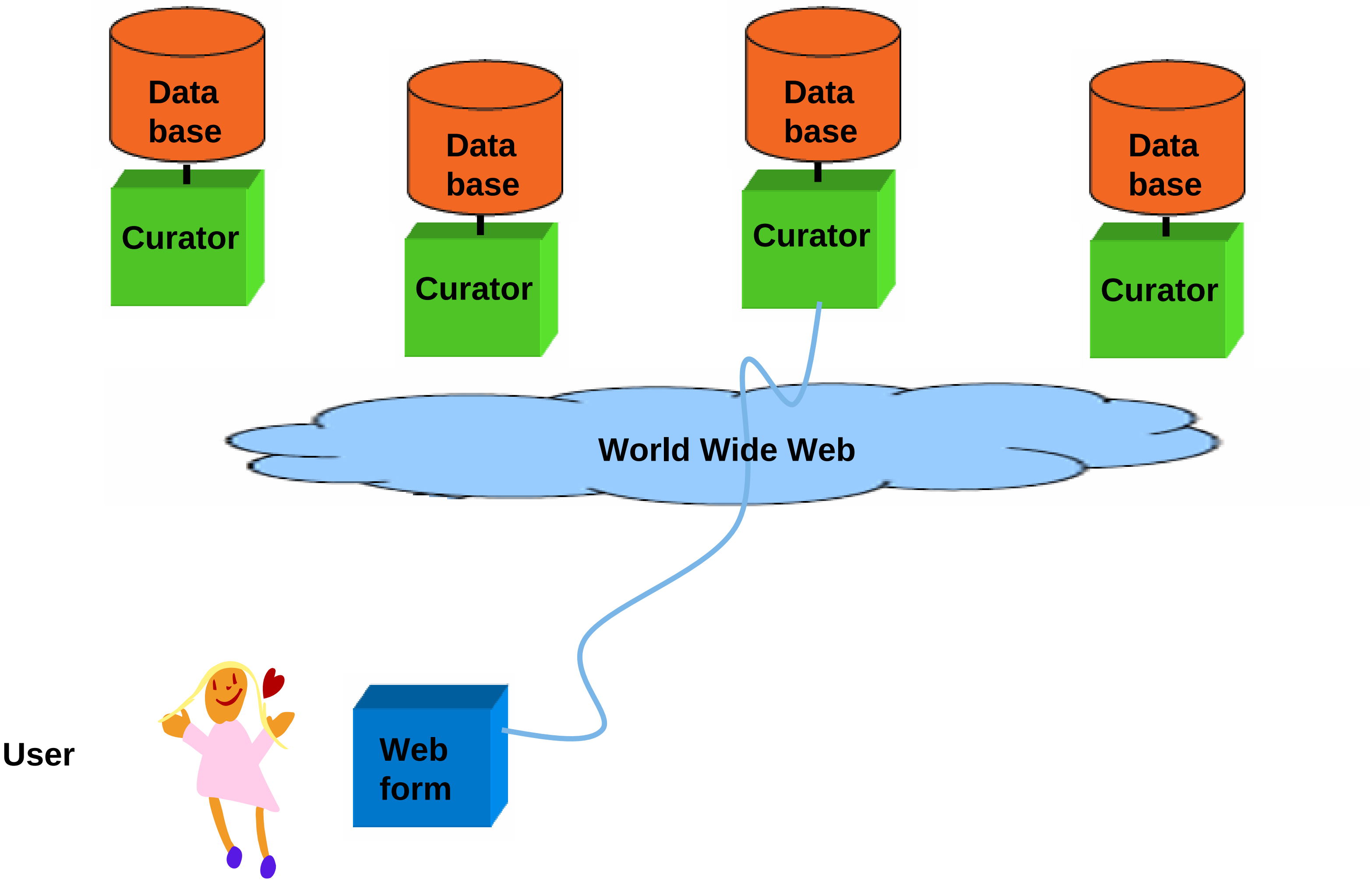
Longer Description (up to 3 lines)

Send Reset

Note:

The data entered will be reviewed by the database administrators before being added to the database; submitted records will usually be available for viewing in one or two working days. Please check back in a few days to confirm that your submission has been added.

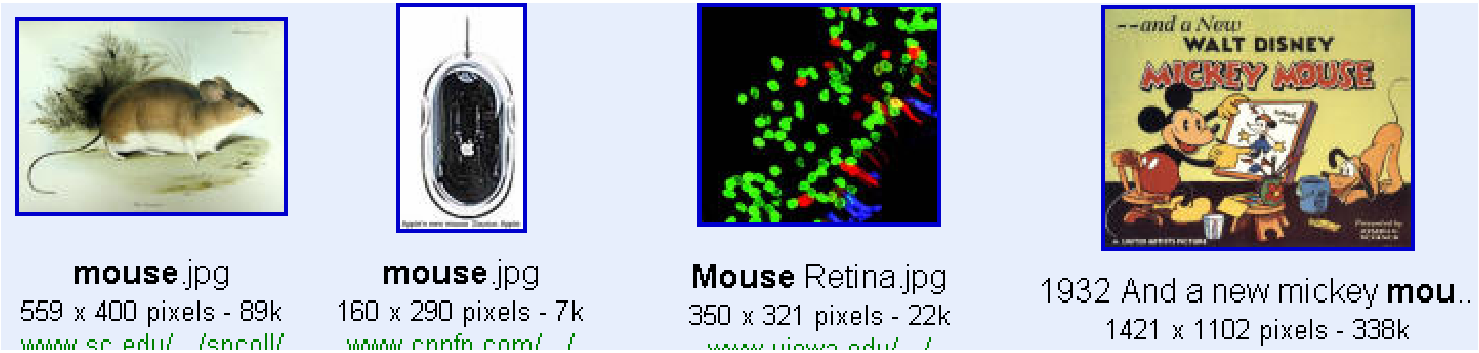
Database submission



Searching online databases

- Searching is by exact keyword matching
- Lack of 'intelligence' and semantic underpinning can lead to frustration

e.g. searching Google Images for "mouse"



mouse.jpg
559 x 400 pixels - 89k
www.sc.edu/.../spcoll/

mouse.jpg
160 x 290 pixels - 7k
www.cnnfn.com/.../

Mouse Retina.jpg
350 x 321 pixels - 22k
www.uiowa.edu/.../

1932 And a new mickey **mou**..
1421 x 1102 pixels - 338k

Unique advantage of semantic searches

The benefits of an ontology-driven database search are potentially enormous . . .

They include the ability to undertake semantically rich searches that can handle

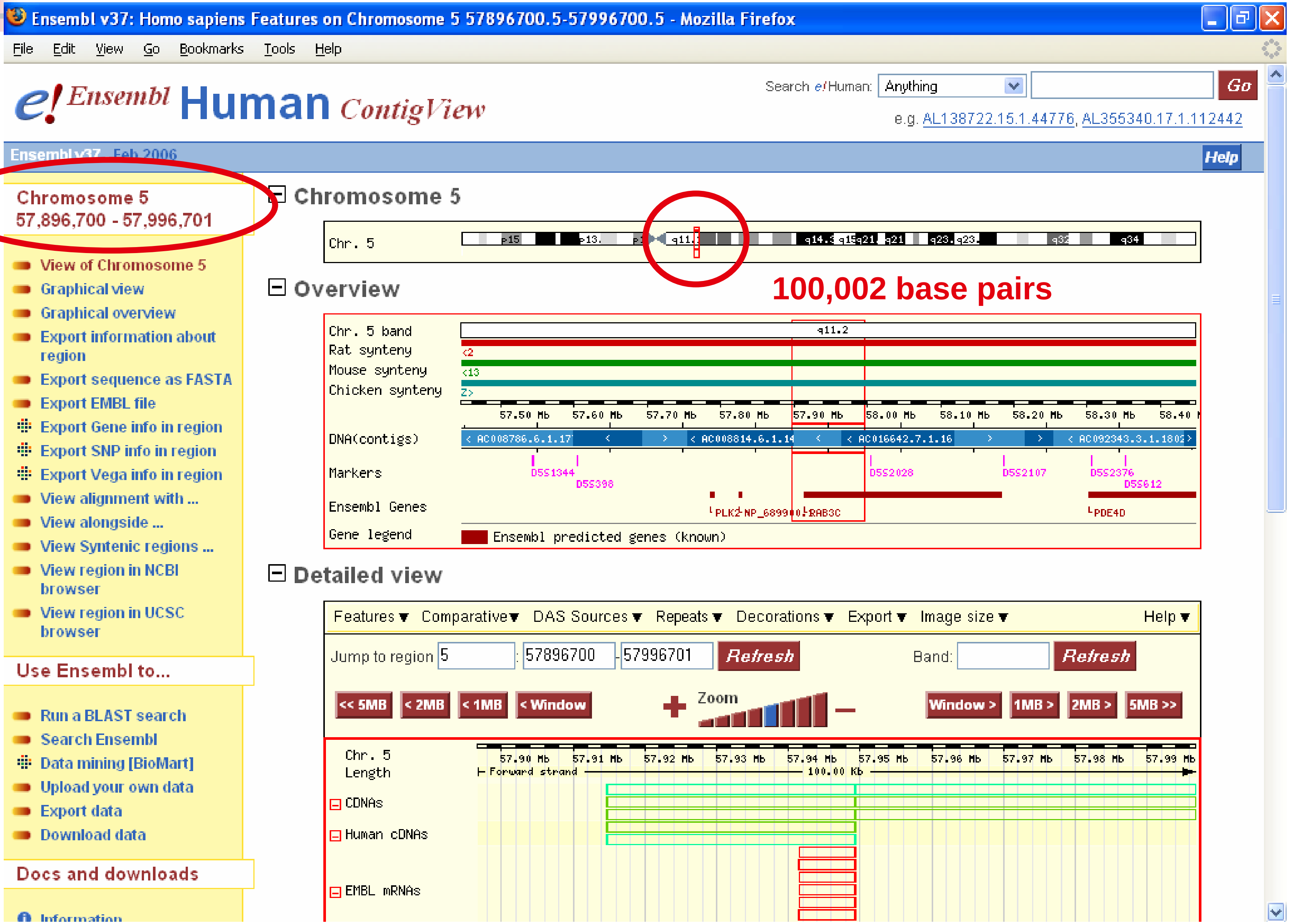
- synonyms (‘mouse’ and ‘*Mus musculus*’)
- homonyms (“mouse” - what does it mean?)
- hierarchies (‘rodent’ and ‘mammal’)
- exclusions (*not* a computer mouse)
- and related terms (‘laboratory animal’ and ‘model species’)

This means that you can search for ‘rodent’ images, even though the image metadata may only contain the terms “mouse” or “rat”

The biological data deluge and its consequences

- Over the last decade, the volume of biological data has grown exponential
  - the current rate of doubling estimated to be every twelve to fifteen months
- Biological research data production is characterised by
  - heterogeneity and lack of central control
  - bottom up data flow from individual labs to bioinformatics databases
- There are more and more independent biological databases
- Laboratory research biologists struggle to keep abreast of relevant information
- However, the availability of sufficient compute power, bandwidth and digital storage to handle the deluge of biological data is not the central issue
- The real problems concern the effective retrieval, analysis and integration of this information, for which semantic annotation and structuring of metadata is vital
- Knowledge creation
  - involves interpretation of new laboratory findings in the light of existing information in the literature and in bioinformatics databases
  - is dependent upon sophisticated tools (e.g. BLAST, ENSEMBL)

Bioinformatics databases can be complex!



Ensembl v37: Homo sapiens Features on Chromosome 5 57896700.5-57996700.5 - Mozilla Firefox

File Edit View Go Bookmarks Tools Help
e! Ensembl Human ContigView
Search e! Human: Anything Go
e.g. AL138722.15.1.44776, AL355340.17.1.112442
Ensembl v37 Feb 2006
Help

Chromosome 5
57,896,700 - 57,996,701

- View of Chromosome 5
- Graphical view
- Graphical overview
- Export information about region
- Export sequence as FASTA
- Export EMBL file
- Export Gene info in region
- Export SNP info in region
- Export Vega info in region
- View alignment with ...
- View alongside ...
- View Syntenic regions ...
- View region in NCBI browser
- View region in UCSC browser

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Export data
- Download data

Docs and downloads

- Information

Chromosome 5

Overview

100,002 base pairs

Gene legend: 🟥 Ensembl predicted genes (known)

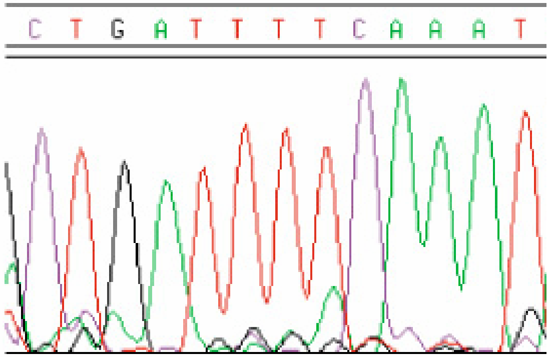
Detailed view

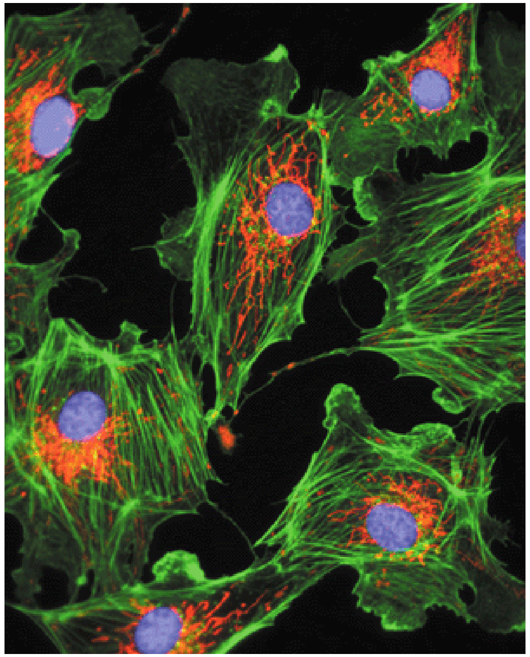
Features ▼ Comparative ▼ DAS Sources ▼ Repeats ▼ Decorations ▼ Export ▼ Image size ▼ Help ▼

Jump to region 5 : 57896700 - 57996701 Refresh Band: Refresh

The dual nature of biological data

- ‘Universal truths’, such as the sequence of a particular gene, or the 3D structure of a specific protein
  - These form bounded data sets
  - The data need only be collected once, and would be the same whoever acquires them
  - Such information is typically published in the public domain
    - It is seen as fundamental research knowledge to which all should have free access
- ‘Particulars’, rather than ‘universals’, for example microscope images of cells:
  - These data form unbounded data sets
  - Data collection will never be complete
  - Such information is not (yet) widely available
    - It is by its nature subject to copyright laws





Storage of data representing ‘universal truths’

- These two types of data need to be stored and published in different ways
- Data representing ‘universal truths’
  - Are of central importance and of finite volume
  - Should be submitted to an appropriate central global database, such as those maintained by the European Bioinformatics InstituteThere is one such database for each data type
  - EMBL, UniProt, PDB, and each of the genome databases
- A note of caution, however:

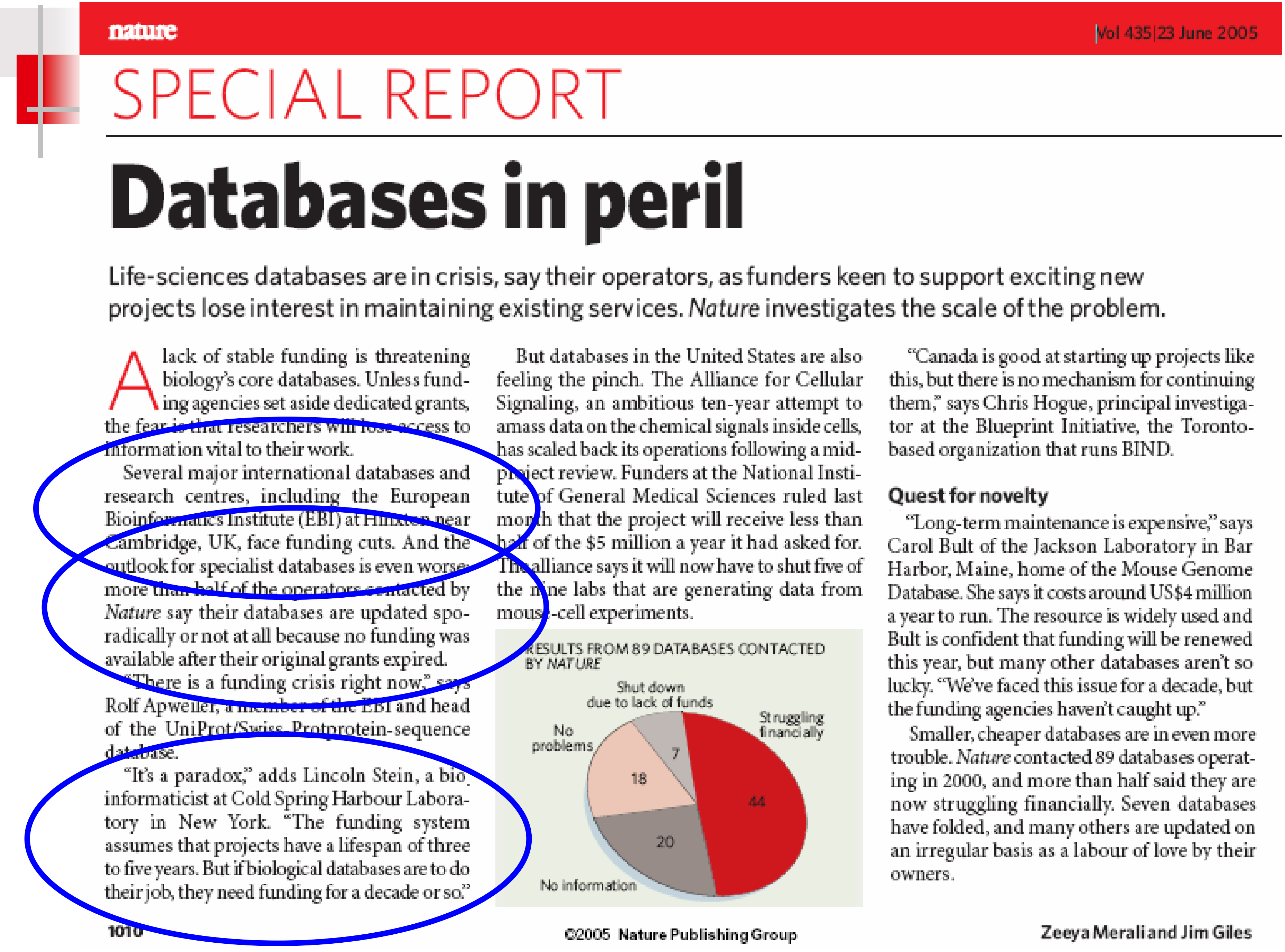
“We believe that attempts to solve [all] the issues of scientific data management by building large, centralised, archival repositories are both dangerous and unworkable.”

(also from *Towards 2020 Science*, Report by Microsoft Research, March 2005)

Storage of data representing ‘particulars’

- Data representing ‘particulars’ form an equally valuable part of the scientific record
- However, they are handled in different ways:
  - They are *NOT* suitable for storage in single global databases
  - Most are never published, although there is an increasing trend to rectify this
  - If they are made public at all, they are housed in distributed specialist databases, typically set up by individual research groups
  - Our new BBSRC-funded *Drosophila* Testis Gene Expression Database (http://www.fly-ted.org) is a good example of such a small specialist database

- However, integration across such highly specialized independent resources is extremely difficult
  - There is a lack of standards for formats and data types
  - Use requires a high level of tacit knowledge, which is different for each resource
  - Database structures and interfaces are subject to modification and upgrading without warning, jeopardising ‘screen-scraping’ methods for automated data harvesting



nature | Vol 435 | 23 June 2005

SPECIAL REPORT

Databases in peril

Life-sciences databases are in crisis, say their operators, as funders keen to support exciting new projects lose interest in maintaining existing services. *Nature* investigates the scale of the problem.

A lack of stable funding is threatening biology's core databases. Unless funding agencies set aside dedicated grants, the fear is that researchers will lose access to information vital to their work.

Several major international databases and research centres, including the European Bioinformatics Institute (EBI) at Hinxton near Cambridge, UK, face funding cuts. And the outlook for specialist databases is even worse: more than half of the operators contacted by *Nature* say their databases are updated sporadically or not at all because no funding was available after their original grants expired.

"There is a funding crisis right now," says Rolf Apweiler, a member of the EBI and head of the UniProt/Swiss-Protprotein-sequence database.

"It's a paradox," adds Lincoln Stein, a bioinformaticist at Cold Spring Harbour Laboratory in New York. "The funding system assumes that projects have a lifespan of three to five years. But if biological databases are to do their job, they need funding for a decade or so."

But databases in the United States are also feeling the pinch. The Alliance for Cellular Signaling, an ambitious ten-year attempt to amass data on the chemical signals inside cells, has scaled back its operations following a mid-project review. Funders at the National Institute of General Medical Sciences ruled last month that the project will receive less than half of the $5 million a year it had asked for. The alliance says it will now have to shut five of the nine labs that are generating data from mouse-cell experiments.

RESULTS FROM 89 DATABASES CONTACTED BY *NATURE*

"Canada is good at starting up projects like this, but there is no mechanism for continuing them," says Chris Hogue, principal investigator at the Blueprint Initiative, the Toronto-based organization that runs BIND.

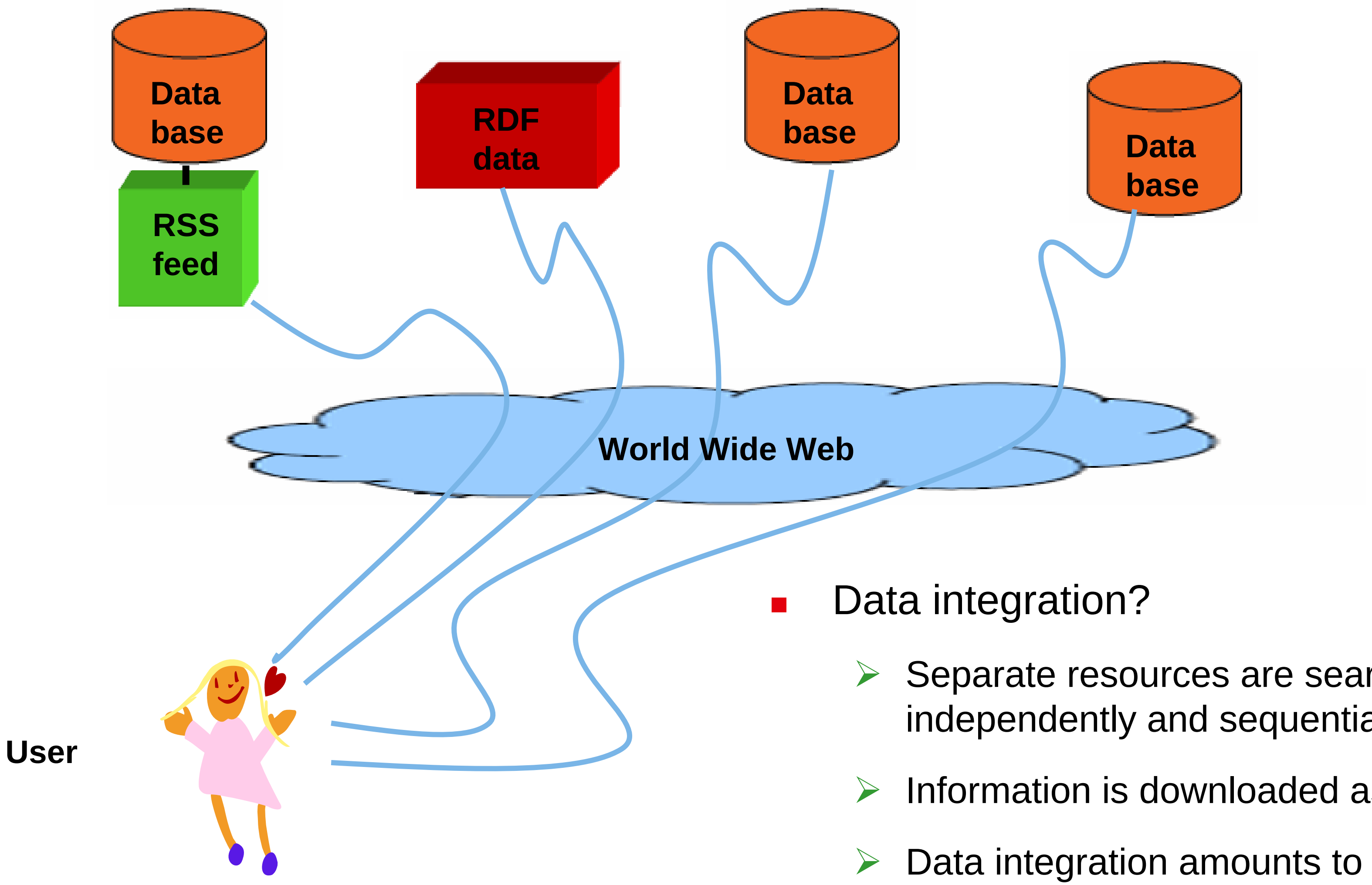
Quest for novelty

"Long-term maintenance is expensive," says Carol Bult of the Jackson Laboratory in Bar Harbor, Maine, home of the Mouse Genome Database. She says it costs around US$4 million a year to run. The resource is widely used and Bult is confident that funding will be renewed this year, but many other databases aren't so lucky. "We've faced this issue for a decade, but the funding agencies haven't caught up."

Smaller, cheaper databases are in even more trouble. *Nature* contacted 89 databases operating in 2000, and more than half said they are now struggling financially. Seven databases have folded, and many others are updated on an irregular basis as a labour of love by their owners.

Zeeya Merali and Jim Giles

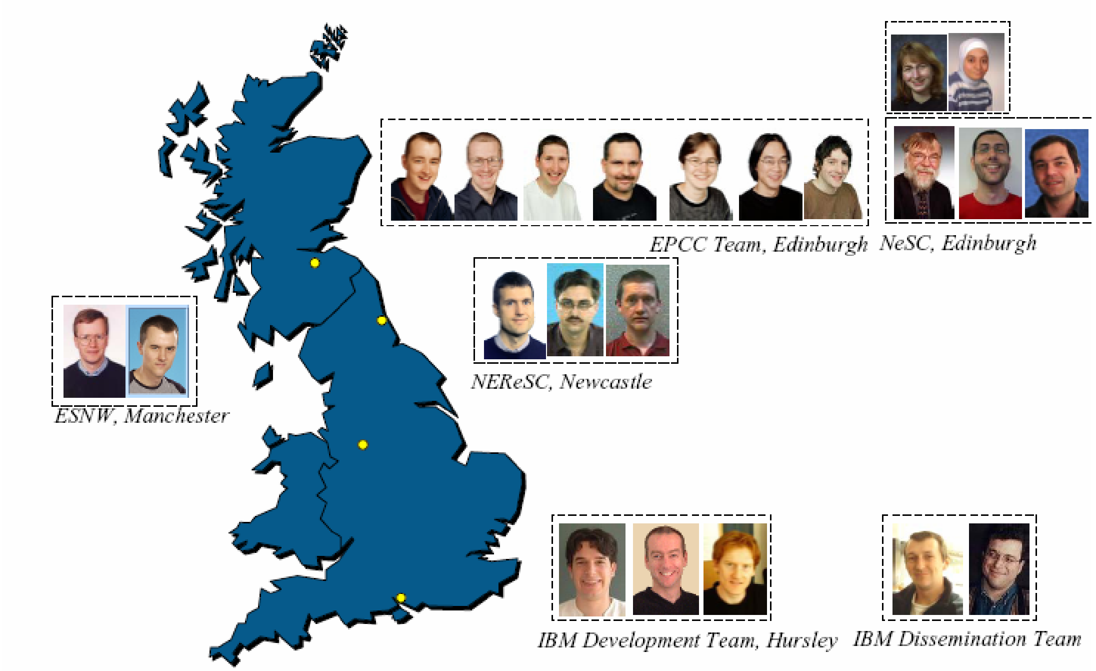
Database integration – the status quo



- Data integration?
  - Separate resources are searched independently and sequentially
  - Information is downloaded as required
  - Data integration amounts to no more that cutting and pasting into a Word document

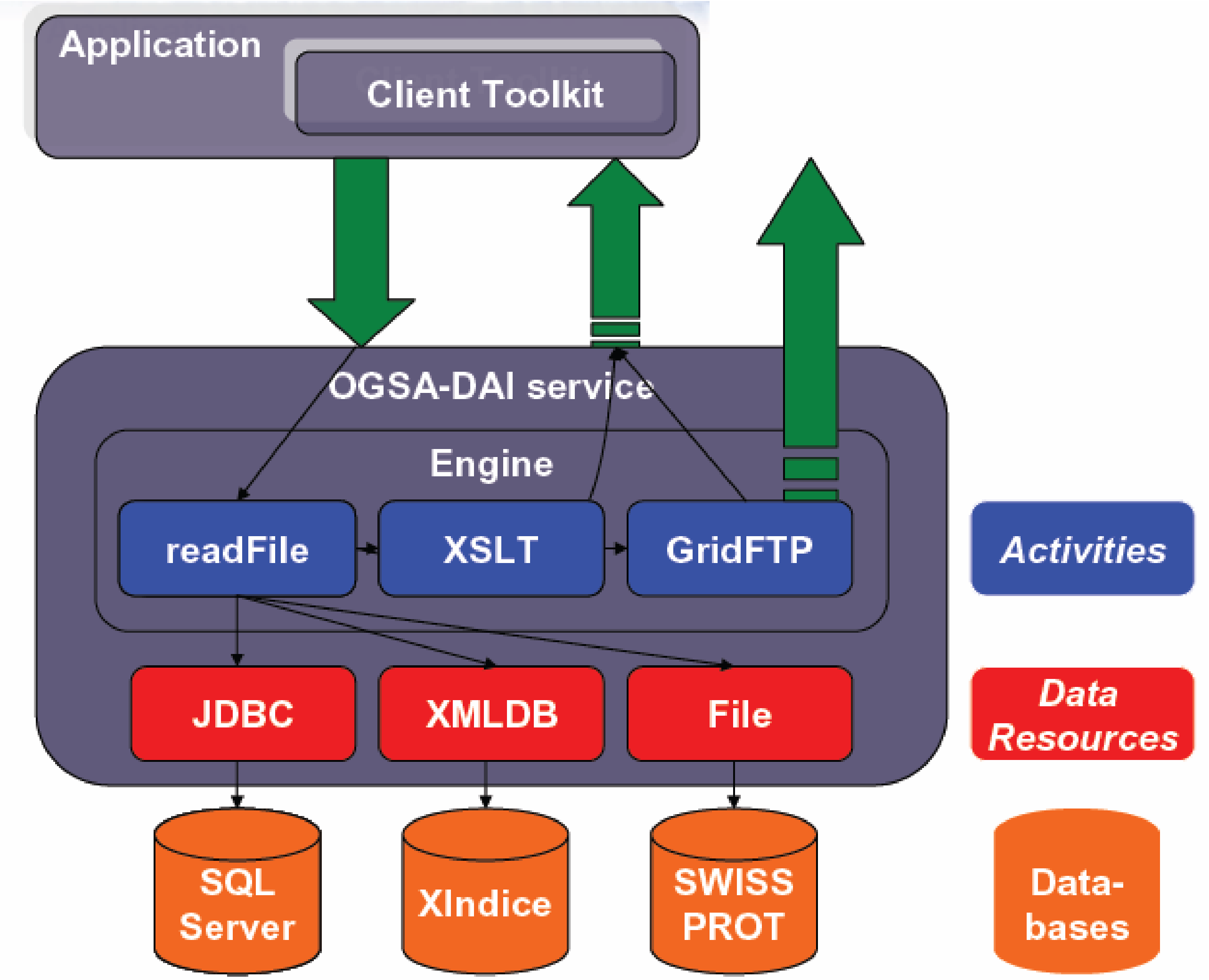
Database integration – the heavyweight approach

- OGSA-DAI
  (Open Grid Services Architecture – Database Access and Integration)
- Mechanism for distributing SQL queries over geographically separate databases
- Heavy investment from UK e-Science budget
- Large development team

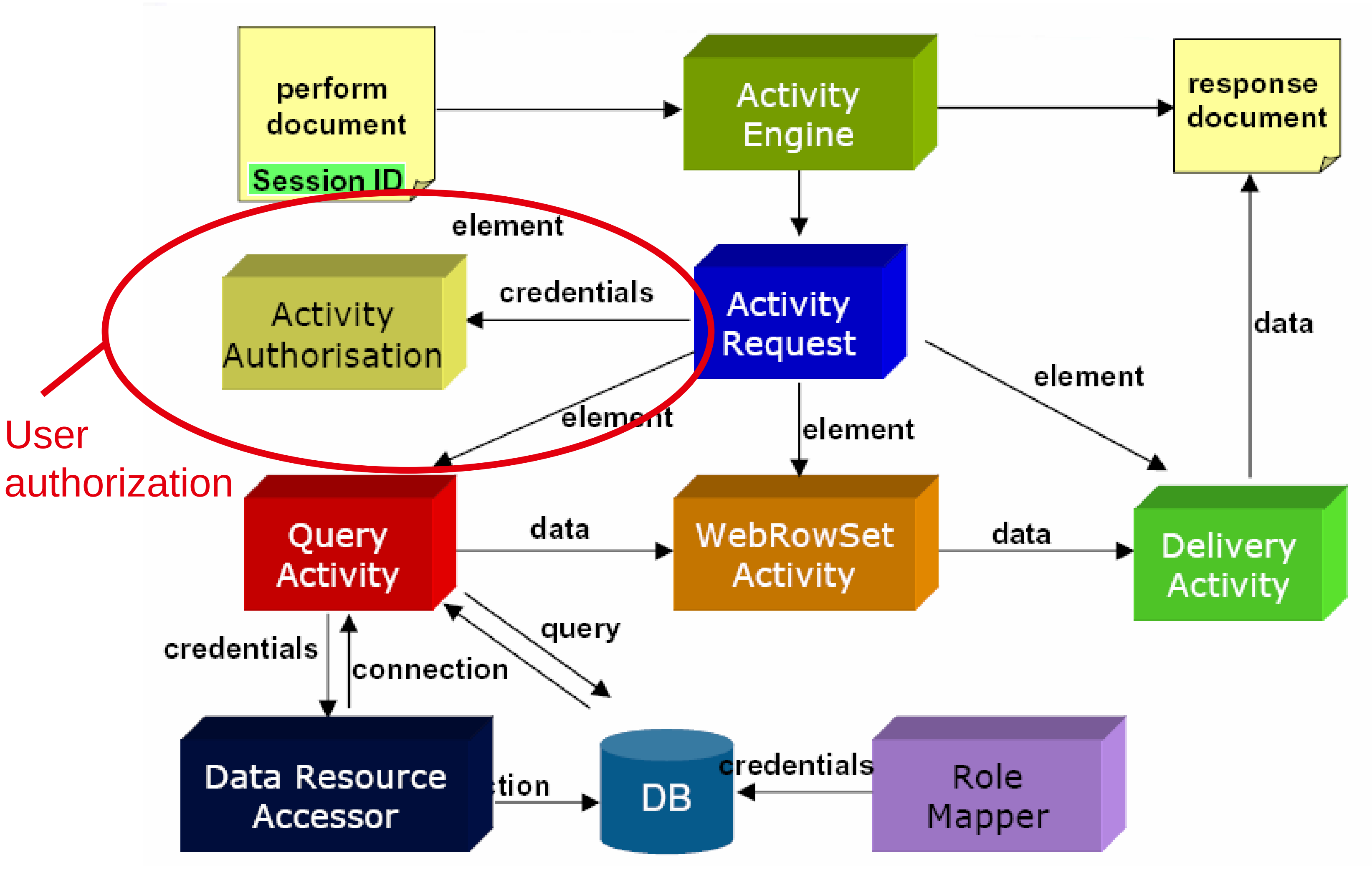


- The following slides are taken from OGSA-DAI Architecture document, 15 Feb 2006, available at http://www.ogsadai.org.uk/documentation/presentations/ggf16/

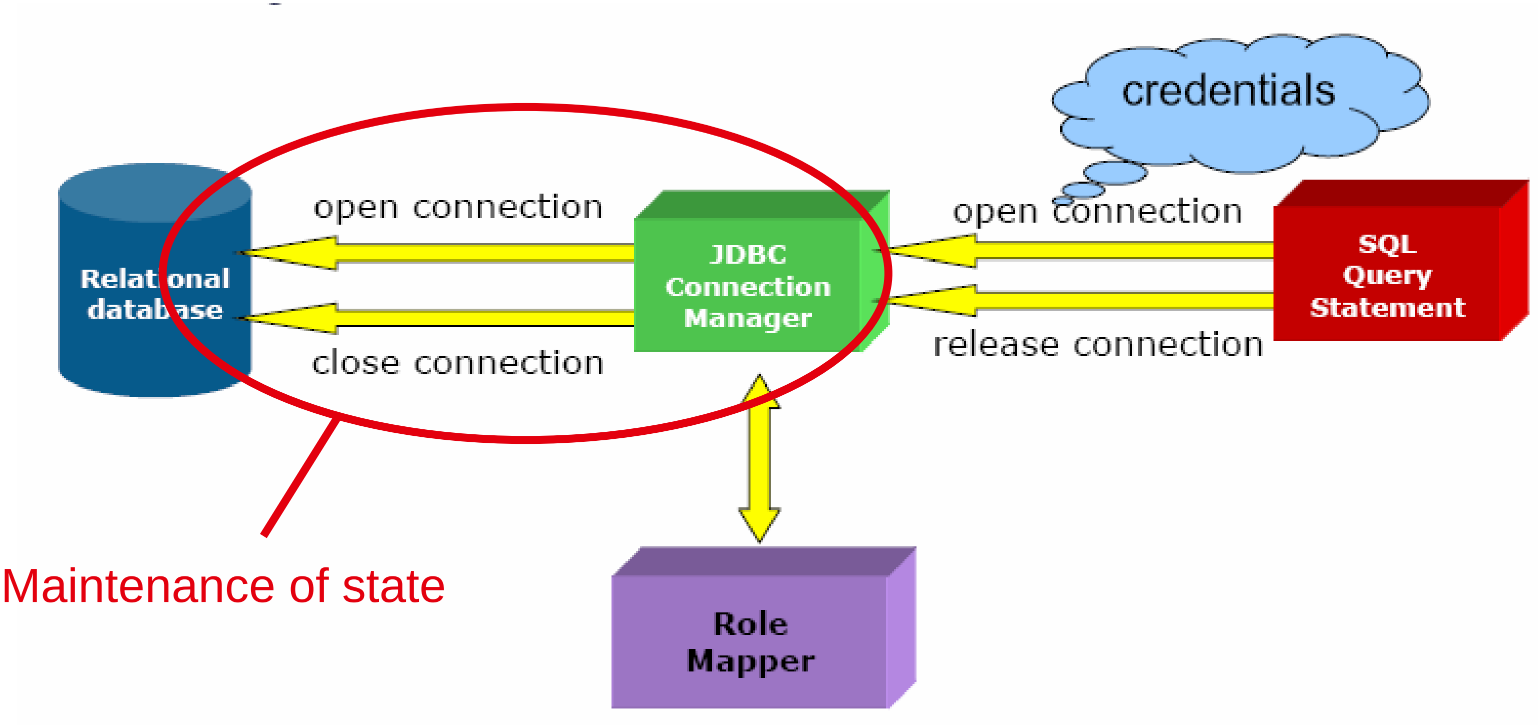
The OGSA-DAI framework



OGSA-DAI data services



OGSA-DAI state management



Maintenance of state

Database integration – the lightweight data web approach

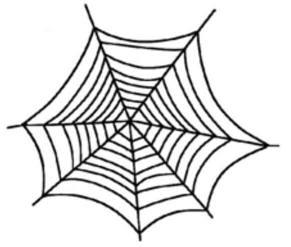
The data web concept

- A data web is a new concept in digital information storage and integration
- The data are *NOT* submitted to a central database, but are simply published in a distributed fashion by the data providers on their own Web servers
- Lightweight semantic web tools are then used to integrate, into a central ontology-enabled registry, metadata describing the distributed data
- All that is required of the data publisher is to make metadata available on his server as RDF, conforming to a particular minimalist data web ontology
- These data can be harvested automatically by searching for the appropriate ontology namespace
- Remember: with RDF, integration comes for free! Thus exporting a database's content as RDF gives immediate inter-operability with other RDF databases

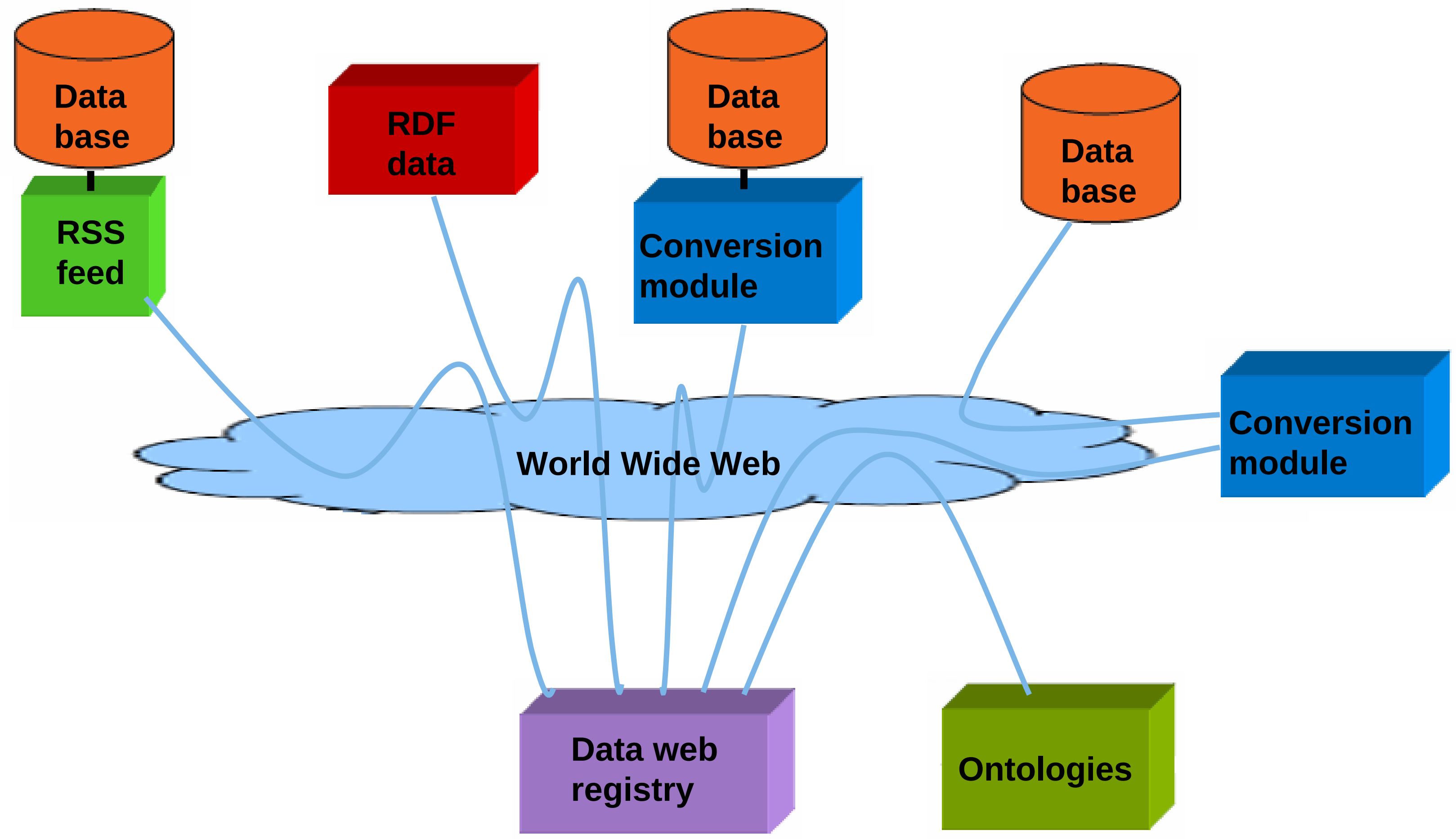
Database access – the lightweight data web approach

Role of the data web registry

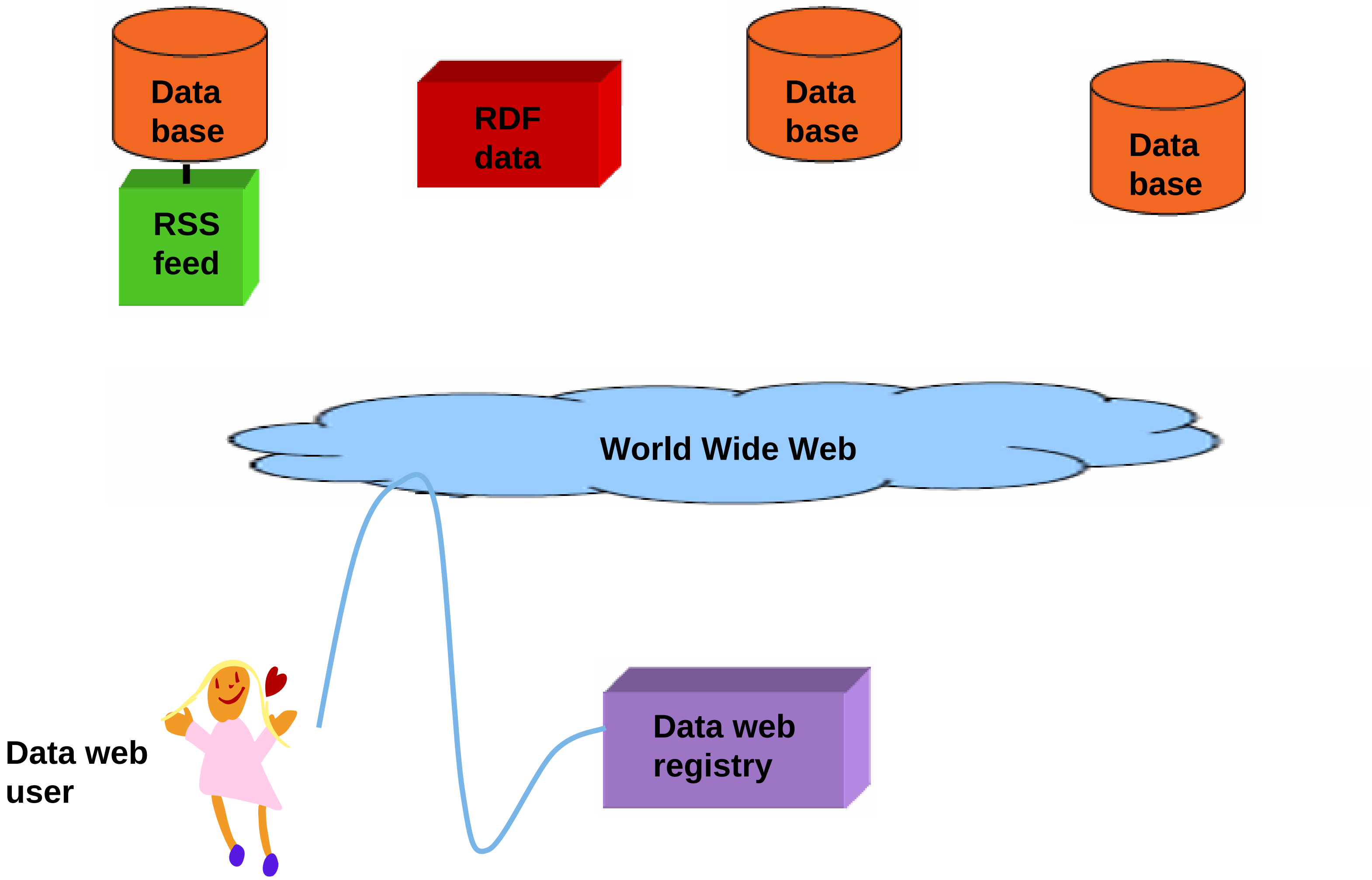
- The data web registry provides an integrated cross-searchable access point to all the data in the data web, thus facilitating access to them and enabling presently impossible meta-research
- The data web registry thus acts for the published data as Google does for conventional Web pages, adding value by providing interoperability and customizable search interfaces, but with a more rigorous semantic underpinning
- The primary data holders benefit by increased web traffic to their sites, while at the same time being able to maintain normal copyright and access control
- The primary data are never owned by the registry, but are freely available for use by other presently unforeseen applications, including novel data integration or analysis services



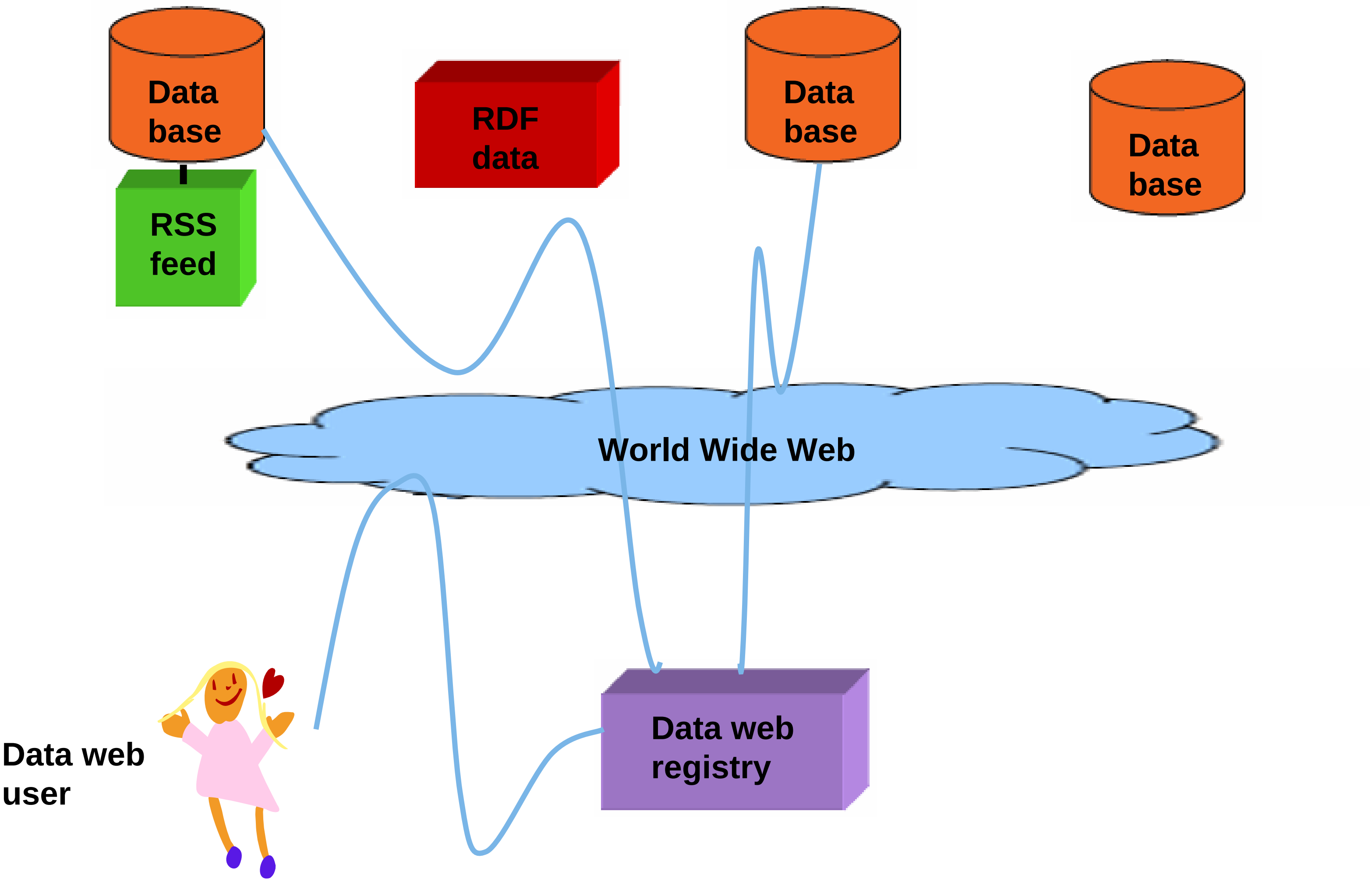
The Data Web Model – data acquisition and indexing



The Data Web Model – user query

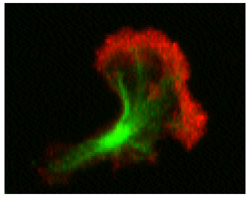


The Data Web Model – user referral



The interpretation of biological images

- Unlike biological sequences and molecular crystal structures, which carry their own internal semantics, images are not self-describing
- The meaning of a biological research image can only be properly understood in the context of the experiment within which it was acquired



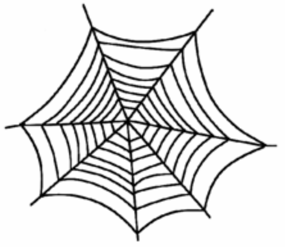
What is it?
A river delta?
Broccoli?

- This requires
  - either that the image be studied in the context of the paper in which it was published
  - or that the image be accompanied by very rich descriptive metadata
- In the BioImageWeb Project, both approaches will be possible:
  - First, the BioImageWeb Registry will be capable of storing rich metadata if available
  - Additionally and more importantly, the BioImageWeb Registry will provide links back to the primary publications on publishers’ sites, where full Methods and Materials descriptions will be available

The BioImageWeb Project purpose

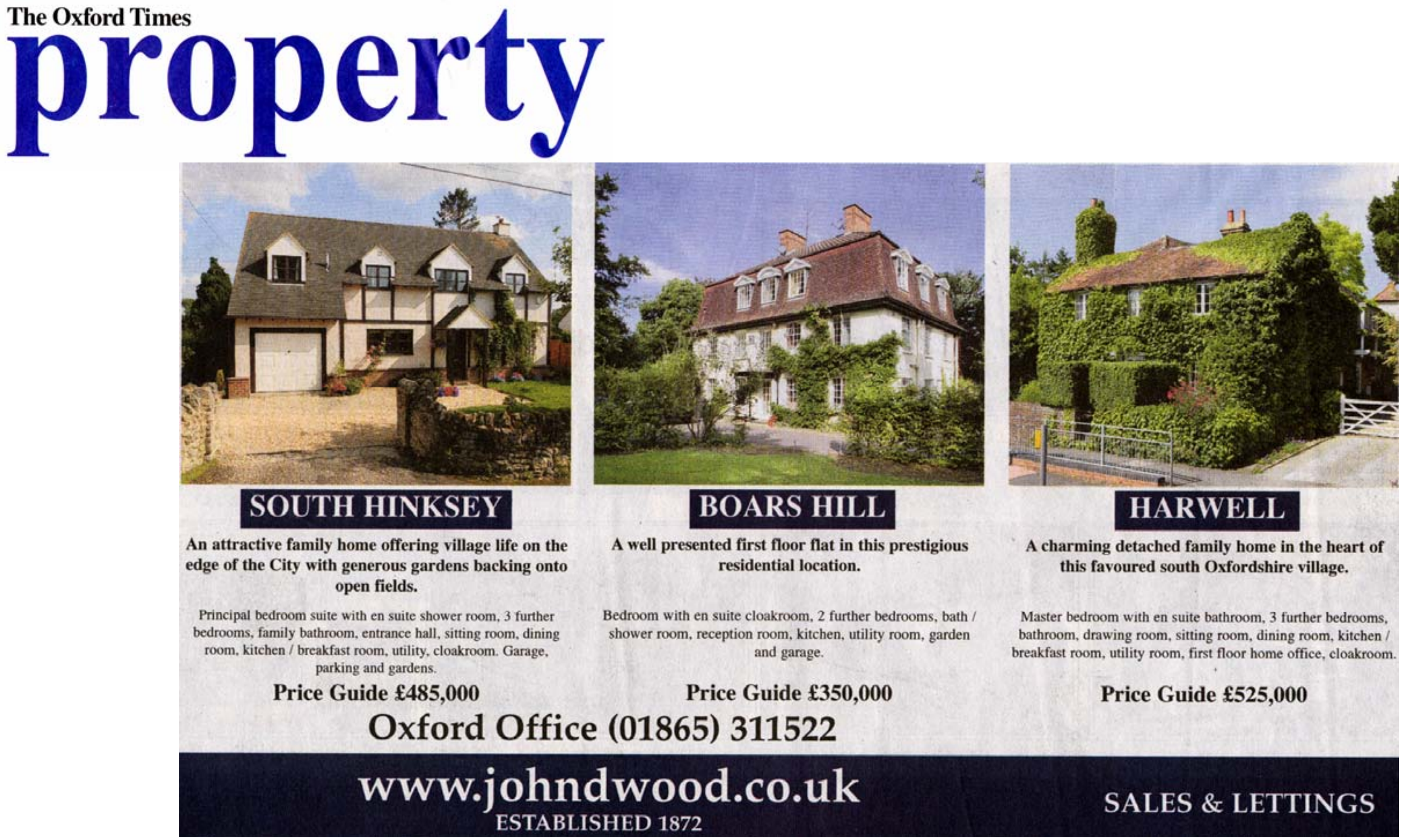
- To integrate and make cross-searchable biological research images held by publishers and institutional repositories, which are currently in isolated data silos
- It should involve minimum effort on the part of the publishers and repository managers, who can use their existing RSS feeds or XML metadata schemas
- We will convert these as necessary to RDF, for example by mapping database tables to the ‘BioImageCore’ ontology, then use D2R for automated conversion
- It requires harvesting of thumbnails and basic metadata describing the images
- We will use our BioImage Database as the metadata registry, from which users will be referred to the original source of the images in their textual context
- Publishers will retain access control to their own journals, and copyright holders will maintain copyright over their image data
- BioImageWeb will enable publishers’ web sites to become a more integral part of day-to-day research, and published images to be used more fully than at present

bioimageweb



The BioImageWeb model – a real world analogy

- The local newspaper property section contains thumbnail images and basic metadata about houses for sale – equivalent to the BioImageWeb Registry
- Users searching this central ‘registry’ pick out what they like, and then . . .



The Oxford Times

property

SOUTH HINKSEY

An attractive family home offering village life on the edge of the City with generous gardens backing onto open fields.

Principal bedroom suite with en suite shower room, 3 further bedrooms, family bathroom, entrance hall, sitting room, dining room, kitchen / breakfast room, utility, cloakroom. Garage, parking and gardens.

Price Guide £485,000

BOARS HILL

A well presented first floor flat in this prestigious residential location.

Bedroom with en suite cloakroom, 2 further bedrooms, bath / shower room, reception room, kitchen, utility room, garden and garage.

Price Guide £350,000

HARWELL

A charming detached family home in the heart of this favoured south Oxfordshire village.

Master bedroom with en suite bathroom, 3 further bedrooms, bathroom, drawing room, sitting room, dining room, kitchen / breakfast room, utility room, first floor home office, cloakroom.

Price Guide £525,000

Oxford Office (01865) 311522

www.johndwood.co.uk

ESTABLISHED 1872

SALES & LETTINGS

. . . go round to the estate agent’s office for full details!

The BioImageWeb Project participants

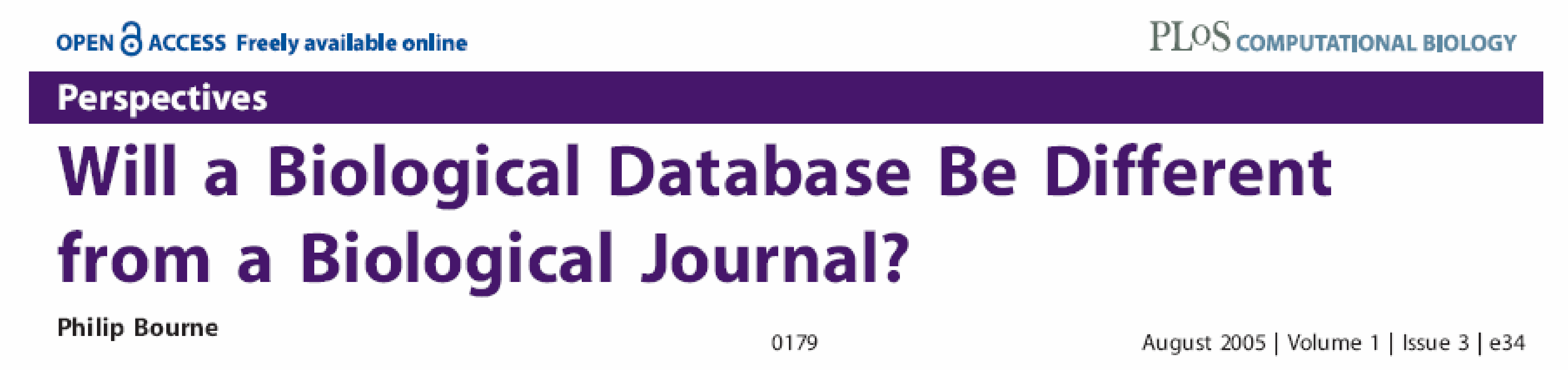
- Image BioInformatics Research Group, University of Oxford
- Leading commercial publishers
  - Nature Publishing Group and Oxford University Press
- Leading Open Access publishers
  - The Public Library of Science and BioMed Central
- University institutional repositories
  - Universities of Cambridge, Imperial College, Oxford and Southampton
- Other stakeholders
  - CrossRef, the Research Information Network, and SPARC Europe
- Professional biologists and academic biological image collections

BioImageWeb advantages and disadvantages

- A data webs such as BioImageWeb has all the advantages of the World Wide Web itself:
  - lack of control
  - freedom and decentralization of publication
  - distributed data
  - a "missing is not broken" Open World philosophy
  - built-in scalability
- However, the fact that the data are coming from selected publishers means that it will not share the Web’s disadvantages of:
  - lack of quality control
  - lack of consistency
- The data web thus overcomes the problems caused by differences in data presentation formats, and makes collating information from multiple web sites possible for machines

Servicing a moving target

- Of course, we must be aware that the concept of an on-line journal is itself changing



OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Perspectives

Will a Biological Database Be Different from a Biological Journal?

Philip Bourne

0179

August 2005 | Volume 1 | Issue 3 | e34

- “Far from limiting themselves to merely linking to databases, scientific journals will in some senses need to become databases. In the longer term, hybrid publications will emerge that combine the strengths of traditional journals with those of databases.”

(again from *Towards 2020 Science*, Report by Microsoft Research, March 2005)

The data deluge . . .

- As the volume of research data accumulates, few if any of us will have the time or the mental capacity to assimilate new data, without first processing them through an ontology or some other similar machine-based organisational aid
- Soon the *only* way to handle the biological data deluge will be through the presuppositional ‘spectacles’ of an ontology
- Does that matter? After all, the ontology is a specification of the accepted paradigm established by the respected leading academics of the day
- In other words, **an ontology fossilized the prejudices of the old farts**

. . . and the paradigm trap

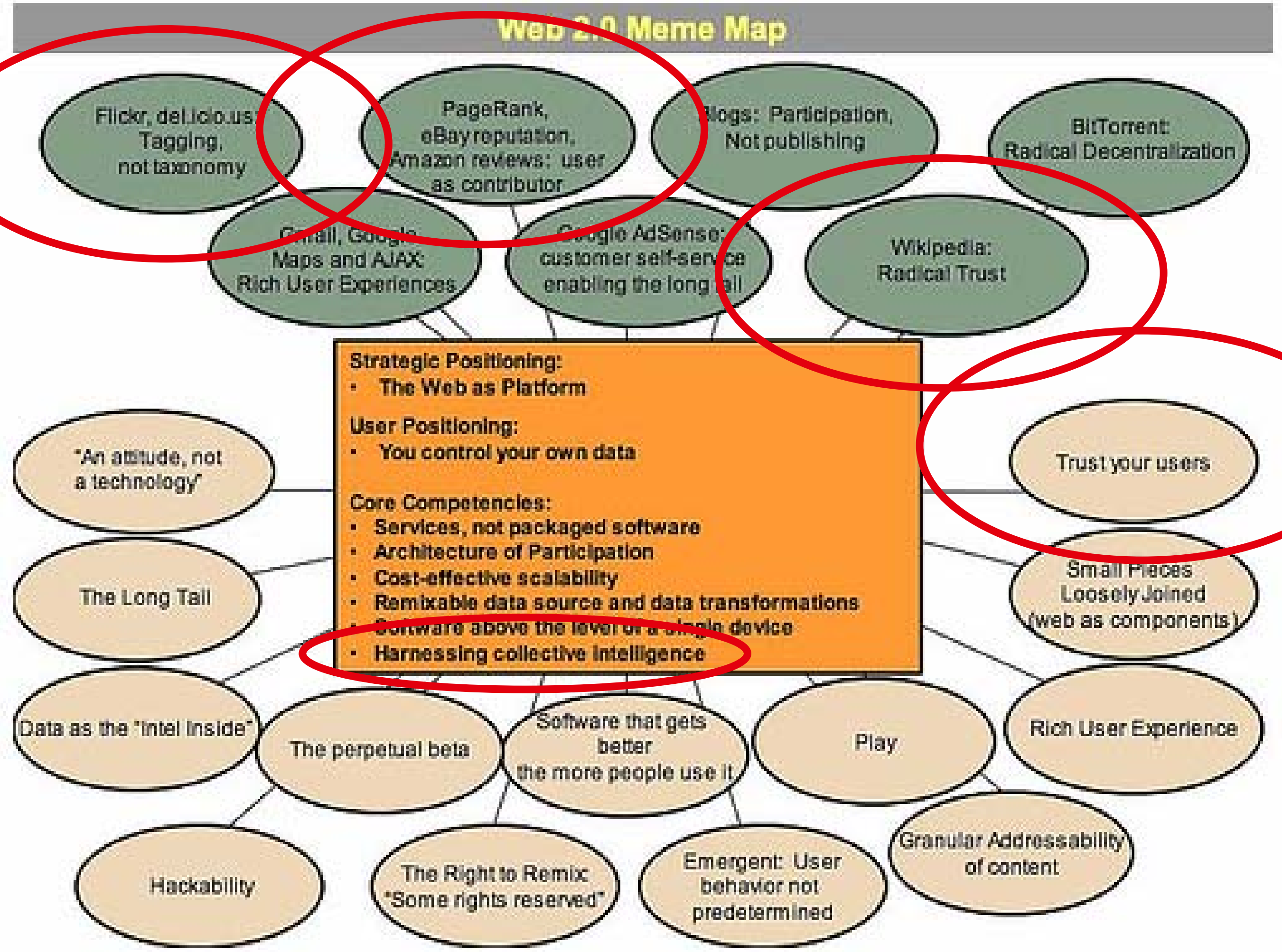
- As first pointed out by Duncan Davidson, there is a danger that information that fits the paradigm will become the ***only*** information ever seen by the user
- This could lead to a blinkered view of the world, which might hamper the process of discovery, prevent the exploration of new and uncharted territory, and inhibit the overthrow of incorrect hypotheses and paradigms

What if Newton had written the ontology for physics?

- The extensive use of defined ontologies could thus make the introduction of radical change even more difficult to achieve
- We need a way for ontologies to evolve with the science, to reflect rather than inhibit paradigm shifts
- And could there be a role for user annotation???

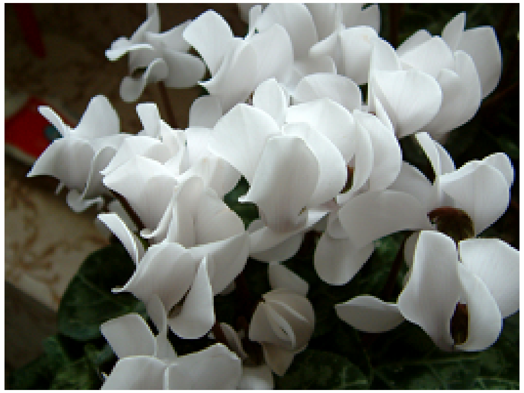
Web 2.0 – usefulness of social tagging

- From Tim O’Reilly’s paper “What is Web 2.0” at http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html


Web 2.0 Meme Map
Flickr, del.icio.us: Tagging, not taxonomy
PageRank, eBay reputation, Amazon reviews: user as contributor
Blogs: Participation, Not publishing
BitTorrent: Radical Decentralization
Gmail, Google Maps and AJAX: Rich User Experiences
Google AdSense: customer self-service enabling the long tail
Wikipedia: Radical Trust
Strategic Positioning:
• The Web as Platform
User Positioning:
• You control your own data
Core Competencies:
• Services, not packaged software
• Architecture of Participation
• Cost-effective scalability
• Remixable data source and data transformations
• Software above the level of a single device
• Harnessing collective intelligence
"An attitude, not a technology"
Trust your users
The Long Tail
Small Pieces Loosely Joined (web as components)
Data as the "Intel Inside"
The perpetual beta
Software that gets better the more people use it
Play
Rich User Experience
Hackability
The Right to Remix: "Some rights reserved"
Emergent: User behavior not predetermined
Granular Addressability of content

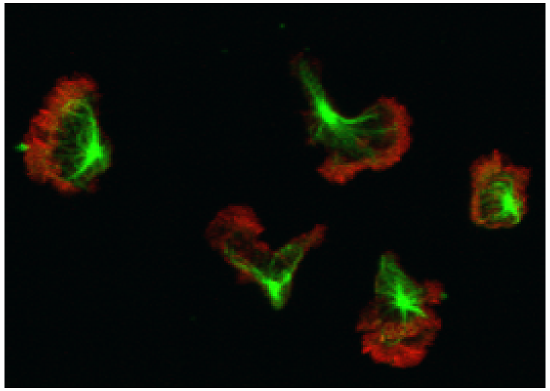
An evaluation of social tagging for scientific image data

1



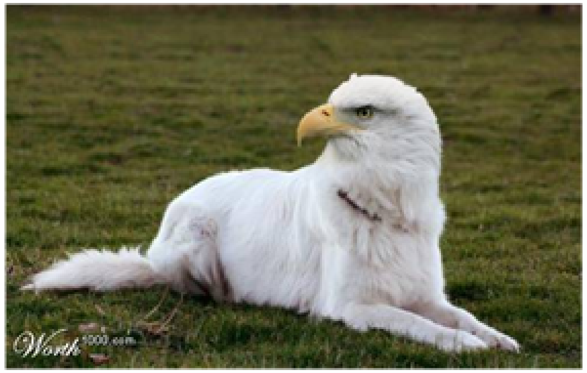
1 . . . .
2 . . . .
3 . . . .
4 . . . .

2



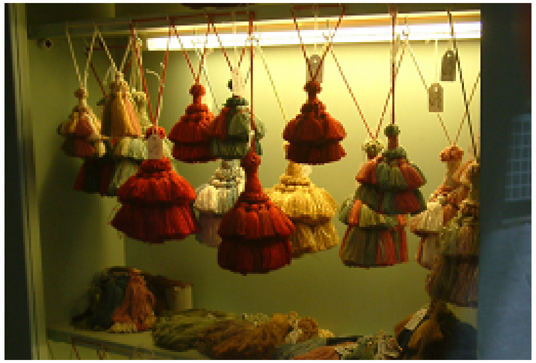
1 . . . .
2 . . . .
3 . . . .
4 . . . .

3



1 . . . .
2 . . . .
3 . . . .
4 . . . .

4



1 . . . .
2 . . . .
3 . . . .
4 . . . .

Acknowledgements

Chris Catton

BioImage Database Development Manager and ImageStore Ontology creator

bioimage

Graham Klyne

Drosophila Testis Gene Expression Database Development Manager

Fly TED