CASIMIR & ENFIN

CASIMIR (FP6: Co-ordination And Sustainability of International Mouse Informatics Resources)

- Addressed the need for scientific integration of mouse genetics and other databases relevant to functional genomics, inside and outside Europe
- Coordinated efforts to standardise data structure, description and exchange
- Examined factors inhibiting deposition of primary data into public databases and sharing of bioresources. IPR and related issues relevant to data policy development and data management.
- Examined models for database funding and financial and scientific sustainability
- 15 publications 2007-2010, incl. Science, Nature commentaries.

Publications

Vol 461|10 September 2009 nature

OPINION

Post-publication sharing of data and tools

Despite existing guidelines on access to data and bioresources, good practice is not widespread. A meeting of mouse researchers in Rome proposes ways to promote a culture of sharing.

Sharing scientific data through publication has long underpinned the cycle of discovery and is the dominant means by which scientists earn credit for their work. More recently, technologies generating very large data sets and novel biological materials have given rise to principles under which communities share data and materials (pre- and post-publication), and to a new sharing infrastructure — large public databases and repositories. Although much attention has been given to practical and ethical guidelines for prepublication data release from large-scale 'community resource projects', summarized in the Bermuda Principles[1] and the Fort Lauderdale report[2], sharing of data and resources from hypothesis-driven research has largely been addressed piecemeal by individual communities, journals and funding agencies.

We report here the efforts of one such community to address issues of particular relevance to the free sharing of data and resources for mouse biology, genetics and functional genomics. Our community has had more than six decades experience with strategies for sharing mice, and more recently for cell lines. When it comes to resource sharing, the two greatest impediments to fully exploiting global research using the mouse as a model organism are the barriers created by material transfer agreements and the underutilization of public mouse repositories.

> "Enforcement of existing policies regarding data and resource deposition is variable."

Community discussion

At a meeting in Rome in May organized by the CASIMIR consortium, a European project examining mouse research infrastructure, participants attempted to establish an agenda for community discussion. This meeting was attended not just by mouse investigators, but by representatives of funding agencies and journals, intellectual-property specialists and sociologists. The resulting Rome Agenda was designed to assist the stakeholders in developing a coordinated and directed approach to the main factors inhibiting free sharing of the fruits of publicly funded mouse research.

Two of the most important shared resources and research outputs in the field are mice and embryonic stem cells. The imperative to share such resources was probably first articulated by the US National Institutes of Health (NIH) in March 1984. Yet even today, numerous unique mouse strains are not made available to the research community despite the existence of publicly funded mouse repositories provided for this purpose (see International Mouse Strain Resource (IMSR) , www.findmice.org). Comparison of the number of knockout mice recorded by the international Mouse Genome Informatics (MGI) database (http://www.informatics.jax.org/) with those deposited in IMSR repositories suggests that currently only 35% are available in this way. This is an encouraging doubling of the percentage available since last assessed in a 2006 NIH survey. To further improve this figure, however, it is important that the sharing ethos is consistently observed by the mouse community and investment in repositories continues to keep pace with the generation of new strains.

Experiences shared at the meeting indicated that enforcement of existing policies regarding data and resource deposition is variable, and that despite increased emphasis on the importance of sharing by journals and funding organizations in recent years, there is evidence that geneticists and genomic researchers are withholding data and research materials with increasing frequency[3]. It is one thing to encourage data deposition and resource sharing through guidelines and policy statements, and quite another to ensure that it happens in practice, as a recent informal survey of proteomics data deposition has revealed[4].

Consequently, although many of the issues discussed in Rome are of specific concern to mouse biology and functional genomics, several have relevance to the wider biological sciences. For example, the issues surrounding licensing and patenting of genetically manipulated mice and embryonic stem cells could apply to many research tools that are generated through hypothesis-driven research. We hope that our experiences and recommendations can inform and stimulate broad discussion in the community as a whole and we ask readers to participate in an online forum to that end (see http://tinyurl.com/mo4gh8).

A strong message from Rome was that funding organizations, journals and researchers need to develop coordinated policies and actions on sharing issues. The Rome Agenda described and summarized here (see "The Rome Agenda," overleaf), represents a challenge to stakeholders to coordinate their efforts to facilitate the ready exchange of data and resources and to share good practices already implemented by some organizations and journals.

Access to publication-associated data

Prepublication data release is comprehensively discussed in an accompanying paper from the Toronto group[5], whose conclusions were broadly supported in Rome. For publication-associated data, the meeting strongly endorsed the recommendations of the National Academy of Sciences UI
out detailed guidel
least the principle
cations are based s
immediately on pub
Currently, fundi
investigators to de
repositories, althou
the consequence th
or not fully exploit
recommended that
journals should ins
onic stem cells be de
tory within a specif
of which still requir
Additionally, funde
itly to cover the costs
ing from projects in
We recommend t
for scientific papers
to access data and r
of the investigation
journals already ha
in this area, insistin
be deposited in pu
all reasonable req
other researchers n
however, heterogen
enforcement; surpri
no written policy o
bioresources or prir
In addition, pap
any other data or m
inating sources. Th
the addition of met
and bioresources[4].
digital object identi
lic repositories, wo
of the literature fo
which is currently

that, unlike diseases such as dengue fever or malaria, but like H1N1, HFMD can be transmitted directly from human to human; no animal reservoir is required. Cases start out with flu-like symptoms that can develop into characteristic sores in the mouth and rashes on the palms and soles of the feet. Infections can also lead to meningitis and swelling of the brain, possibly resulting in long-term neurological damage [L.-Y. Chang *et al. N. Engl. J. Med.* **356**, 1226–1234; 2007] or even death. The vast majority of reported cases and almost all of the serious cases are in patients under 5 years old.

In 2008, more than 500,000 HFMD infections were reported in China, and 200 people died. This year, as the disease moves through its peak season, similar numbers look possible. There is currently no vaccine to prevent it and no drug to cure it. The best safeguard is good hygiene: HFMD transmission is mainly via saliva and faeces. But even Singapore, renowned for its cleanliness, has had 8,896 cases this year, and had had even more by this time last year. There have also been cases this year in Taiwan, Hong Kong and South Korea.

Most of the serious HFMD cases seem to be caused by enterovirus 71 (EV71), first isolated in California in 1969. EV71, in turn, is related to poliovirus — a fact that has some experts concerned. EV71's recent pattern, with a few isolated outbreaks that are building in frequency, is similar to that seen f
that killed or crippled t
Modlin *N. Engl. J. Med.*
certain that EV71 will
to have become increa
because of genetic chan

Thus the balancing a
organizations are to be
to H1N1 (see page 756
by EV71, and its poter
they should also be m
counter that virus. An
Beijing earlier this yea

Yet that good start is
the importance of EV7
on the WHO website.
seem to be going to in
in China still misclas
symptoms change, an
identify it. More leade
much more research o

The sharing principle

Journals and funders must insist that genetically modified mouse strains are fully available.

Back in 1996, human-genome scientists signed up to the Bermuda Agreement to share their data without delay. Since then, the sharing principle has entered the mainstream — it now applies to all genomic data generated using public funding, as well as to all the relevant resources cited in publications.

But this principle is not universally observed for genetically modified mice, designed as vital resources in the quest to unpick basic biological mechanisms or to model human disease. The size of the problem is unclear, but existing surveys, combined with extensive anecdotal evidence, suggest it is substantial. In April 2006, for example, scientists at the US National Institutes of Health found that nearly 1,000 unique mice strains had been created, yet barely 700 had been placed in a repository.

Some scientists say they do not have the time nor money to breed and distribute their mice, or even to send the animals to publicly funded mouse repositories such as the European Mouse Mutant Archive in Europe, the Jackson Laboratory in Bar Harbor, Maine, and coordinate and sustai
ground documents at
brought together repr
ers and the mouse rep
Australasia. They conc
resolution — not least
cally generate mouse l
generate thousands of

To solve the problem
must take a tougher li
few actually requiring
ies wherever possible
simply 'encourage' th
lications freely availal
mice be deposited in r
such cajoling terms as
resources, and rarely p

Journals should no
repositories as a conc
should require reposit
that are likely to gener
should be reserved for
grants should refer to t

nature
www.nature.com/nature Vol 461 | Issue no. 7261 | 10 September 2009

Data's shameful neglect

Research cannot flourish if data are not preserved and made accessible. All concerned must act accordingly.

More and more often these days, a research project's success is measured not just by the publications it produces, but also by the data it makes available to the wider community. Pioneering archives such as GenBank have demonstrated just how powerful such legacy data sets can be for generating new discoveries — especially when data are combined from many laboratories and analysed in ways that the original researchers could not have anticipated.

All but a handful of disciplines still lack the technical, institutional and cultural frameworks required to support such open data access (see pages 168 and 171) — leading to a scandalous shortfall in the sharing of data by researchers (see page 160). This deficiency urgently needs to be addressed by funders, universities and the researchers themselves.

Research funding agencies need to recognize that preservation of and access to digital data are central to their mission, and need to be supported accordingly. Organizations in the United Kingdom, for instance, have made a good start. The Joint Information Systems Committee, established by the seven UK research councils in 1993, has made data-sharing a priority, and has helped to establish a Digital Curation Centre, headquartered at the University of Edinburgh, to be a national focus for research and development into data issues. Other European agencies have also pursued initiatives.

The United States, by contrast, is playing catch-up. Since 2005, a 29-member Interagency Working Group on Digital Data has been trying to get US funding agencies to develop plans for how they will support data archiving — and just as importantly, to develop policies on what data should and should not be preserved, and what exceptions should be made for reasons such as patient privacy. Some agencies have taken the lead in doing so; many more are hanging back. They should all being moving forwards vigorously.

What is more, funding agencies and researchers alike must ensure that they support not only the hardware needed to store the data, but also the software that will help investigators to do this. One important facet is metadata management software: tools that streamline the tedious process of annotating data with a description of what the bits mean, which instrument collected them, which algorithms have been used to process them and so on — information that is essential if other scientists are to reuse the data effectively.

Also necessary, especially in an era when data can be mixed and combined in unanticipated ways, is software that can keep track of which pieces of data came from whom. Such systems are essential if tenure and promotion committees are to give credit — as they should — to candidates' track-record of data contribution.

> "Data management should be woven into every course in science."

Who should host these data? Agencies and the research community together need to create the digital equivalent of libraries: institutions that can take responsibility for preserving digital data and making them accessible over the long term. The university research libraries themselves are obvious candidates to assume this role. But whoever takes it on, data preservation will require robust, long-term funding. One potentially helpful initiative is the US National Science Foundation's DataNet programme, in which researchers are exploring financial mechanisms such as subscription services and membership fees.

Finally, universities and individual disciplines need to undertake a vigorous programme of education and outreach about data. Consider, for example, that most university science students get a reasonably good grounding in statistics. But their studies rarely include anything about information management — a discipline that encompasses the entire life cycle of data, from how they are acquired and stored to how they are organized, retrieved and maintained over time. That needs to change: data management should be woven into every course in science, as one of the foundations of knowledge. ■

A step too far?

The Obama administration must fund human space flight adequately, or stop speaking of 'exploration'.

After the space shuttle *Columbia* burned up during re-entry into Earth's atmosphere in 2003, the board that was convened to investigate the disaster looked beyond its technical causes to NASA's organizational malaise. For decades, the board pointed out, the shuttle programme had been trying to do too much with too little money. NASA desperately needed a clearer vision and a better-defined mission for human space flight.

The next year, then-President George W. Bush attempted to supply that vision with a new long-term goal: first send astronauts to build a base on the Moon, then send them to Mars. This idea immediately set off a debate that is still continuing, in which sceptics ask whether there is any point in returning to the Moon nearly half a century after the first landings. Why not go to Mars directly, or visit near-Earth asteroids, or send people to service telescopes in the deep space beyond Earth?

Yet that debate is both counter-productive — a new set of rockets could go to all of these places — and moot, because Bush's vision never attracted the hoped-for budget increases. Indeed, a blue-riband commission reporting to US President Barack Obama this week (see page 153) finds the organizational malaise unchanged: NASA is still doing too much with too little. Without more money, the agency won't be sending people anywhere beyond the International Space Station, which resides in low Earth orbit only 350 kilometres up. And even the ability to do that is in question: Ares I, the US rocket that would return

Disease Models & Mechanisms 2, 423-425 (2009) doi:10.1242/dmm.004259
Published by The Company of Biologists 2009

Resources, repositories and rewards

Vivian Siegel, Editor-in-Chief

By establishing the Resource Article section and stronger policies for materials-sharing and citation, we hope to encourage and properly reward the development and sharing of resources, thereby accelerating research using model organisms to advance human health

Scientific communities prosper on shared information and material, which allow for both confirmation and advancement of research. Journals, as a nexus of communication and as enablers of scientific work, generally, and rightfully, insist that the information obtained and the materials created in the course of doing published work be made available for future research use – the publishing quid pro quo. Yet, while researchers thrive on the acclaim afforded by publishing, many resist sharing their results and reagents – this is perhaps to their advantage in the short term, but to the detriment of expedient scientific progress in the long term. After several frustrated attempts to obtain a published reagent, scientists often end up either making the reagent themselves or changing projects; each year, one in nine scientists abandons a project because of a denied request for 'research input' (material or data) (Walsh et al., 2005; www.casimir.org.uk/storyfiles/66.0.09_00_Walsh.pdf).

For reagents that are simple and cheap to recreate (such as DNA clones), it is often faster to make the reagent than to ask for it. But for those materials that are expensive and time consuming to create, such as the model organisms that are central to the interests of this journal, we need more insistent policies, supportive infrastructures and rewards to ensure the timely sharing of research materials.

Although journals have had materials-sharing policies in place for some time, these policies have largely failed. In 2001, the National Academy of Sciences convened a committee to propose an explicit solution to the problem (www.nap.edu/catalog.php?record_id=10613). From this came UPSIDE: 'Uniform Principle for Sharing

nature Vol 461|10 September 2009

OPINION

Prepublication data sharing

Rapid release of prepublication data has served the field of genomics well. Attendees at a workshop in Toronto recommend extending the practice to other biological data sets.

Open discussion of ideas and full disclosure of supporting facts are the bedrock for scientific discourse and new developments. Traditionally, published papers combine the salient ideas and the supporting facts in a single discrete 'package'. With the advent of methods for large-scale and high-throughput data analyses, the generation and transmission of the underlying facts are often replaced by an electronic process that involves sending information to and from scientific databases. For such data-intensive projects, the standard requirement is that all relevant data must be made available on a publicly accessible website at the time of a paper's publication[1].

One of the lessons from the Human Genome Project (HGP) was the recognition that making data broadly available before publication can be profoundly valuable to the scientific enterprise and lead to public benefits. This is particularly the case when there is a community of scientists that can productively use the data quickly — beyond what the data producers could do themselves in a similar time period, and sometimes for scientific purposes outside the original goals of the project.

The principles for rapid release of genome-sequence data from the HGP were formulated at a meeting held in Bermuda in 1996; these were then implemented by several funding agencies. In exchange for 'early release' of their data, the international sequencing centres retained the right to be the first to describe and analyse their complete data sets in peer-reviewed publications. The draft human genome sequence[2] was the highest profile data set rapidly released before publication, usually within 24 hours of generation. This experience demonstrated that the broad and early availability of sequence data greatly benefited life sciences research by leading to many new insights and discoveries[3], including new information on 30 disease genes published prior to the draft sequence.

At a time when advances in DNA sequencing technologies mean that many more laboratories can produce massive data sets, and when an ever-growing number of fields (beyond genome sequencing) are grappling with their own data-sharing policies, a Data Release Workshop was convened in Toronto in May 2009 by Genome Canada and other funding agencies. The meeting brought together a diverse and multinational group of scientists, ethicists, lawyers, journal editors and funding representatives. The goal was to reaffirm and refine, where needed, the policies related to the early release of genomic data, and to extend, if possible, similar data release policies to other types of large biological data sets — whether from proteomics, biobanking or metabolite research.

Building on the past

By design, the Toronto meeting continued policy discussions from previous meetings, in particular the Bermuda meetings (1996, 1997 and 1998)[1,2] and the 2003 Fort Lauderdale meeting, which recommended that rapid prepublication release be applied to other data sets whose primary utility was a resource for the scientific community, and also established the responsibilities of the resource producers, resource users, and the funding agencies[4]. A similar 2008 Amsterdam meeting extended the principle of rapid data release to proteomics data[5]. Although the recommendations of these earlier meetings can apply to many genomics and proteomics projects, many outside the major sequencing centres and funding agencies remain unaware of the details of these policies, and so one goal of the Toronto meeting was to reaffirm the existing principles for early data release with a wider group of stakeholders.

In Toronto, attendees endorsed the value of rapid prepublication data release for large reference data sets in biology and medicine that have broad utility and agreed that prepublication data release should go beyond genomics and proteomics studies to other data sets — including chemical structure, metabolomic and RNAi data sets, and to annotated clinical resources (cohorts, tissue banks and case-control studies). In each of these domains, there are diverse data types and study designs, ranging from the large-scale 'community resource projects' first identified at Fort Lauderdale (for which meeting participants endorsed prepublication data release) to investigator-led hypothesis-testing projects (for which the minimum standard should be the release of generated data at the time of publication).

Several issues discussed at previous data-

EXAMPLES OF PREPUBLICATION DATA-RELEASE GUIDELINES

Project type	Pre-publication data release recommended	Pre-publication data release optional
Genome sequencing	Whole-genome or mRNA sequence(s) of a reference organism or tissue	Sequences from a few loci for cross-species comparisons in a limited number of samples
Polymorphism discovery	Catalogue of variants from genomic and/or transcriptomic samples in one or more populations	Variants in a gene, a gene family, or a genomic region in selected pedigrees or populations
Genetic association studies	Genomewide association analysis of thousands of samples	Genotyping of selected gene candidates
Somatic mutation discovery	Catalogue of somatic mutations in exomes or genomes of tumour and non-tumour samples	Somatic mutations of a specific locus or limited set of genomic regions
Microbiome studies	Whole-genome sequence of microbial communities in different environments	Sequencing of target locus in a limited number of microbiome samples
RNA profiling	Whole-genome expression profiles from a large panel of reference samples	Whole-genome expression profiles of a perturbed biological system(s)
Proteomic studies	Mass spectrometry data sets from large panels of normal and disease tissues	Mass spectrometry data sets from a well-defined and limited set of tissues
Metabolomic studies	Catalogue of metabolites in one or more tissues of an organism	Analogue of metabolites induced in a perturbed biological system(s)
RNAi or chemical library screen	Large-scale screen of a cell line or organism analysed for standard phenotypes	Focused screens used to validate a hypothetical gene network
3D-structure elucidation	Large-scale cataloguing of 3D structures of proteins or compounds	3D structure of a synthetic protein or compound elucidated in the context of a focused project

ENFIN (*E*xperimental *N*etwork for *F*unctional *In*tegration)

About

- **ENFIN is a virtual institute, formed to enable systems-level integration of experimental results.**

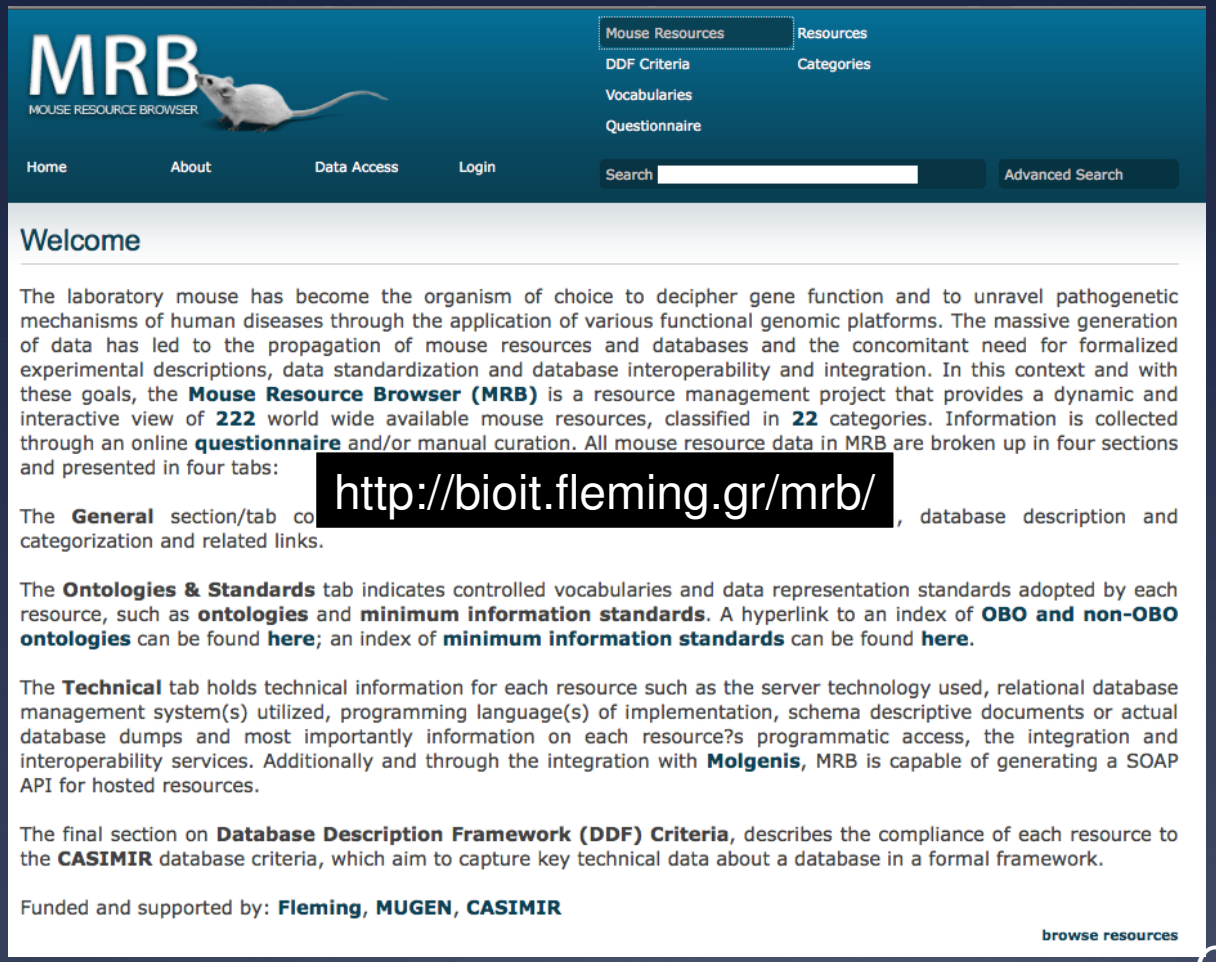
Objectives

- To develop a shared approach between traditionally dry and traditionally wet researchers in the area of systems-level interpretation of experimental results
- To develop a distributed computational platform this integration and analysis of experimental data
- To directly prove that such an approach has scientific value
- To encourage and participate in the critical assessment of systems-level approaches
- To disseminate knowledge and techniques to other academic researchers worldwide
- To disseminate knowledge and techniques to commercial researchers, in particular European SMEs
- To train young European researchers from a variety of backgrounds in system-level informatics techniques.
- All these objectives taken together will enhance both the European Research area and the competiveness of Europe.

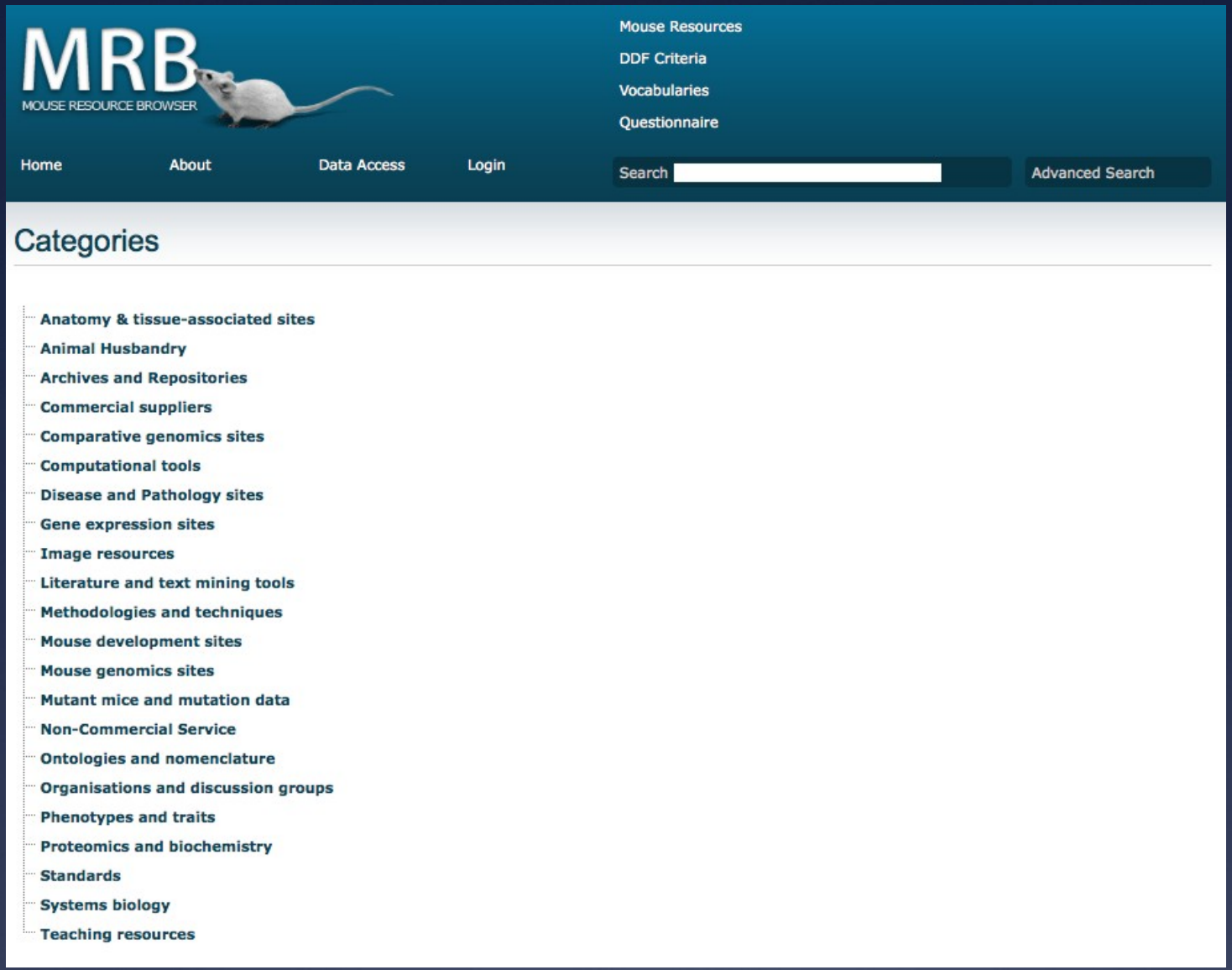
Origins of DDF

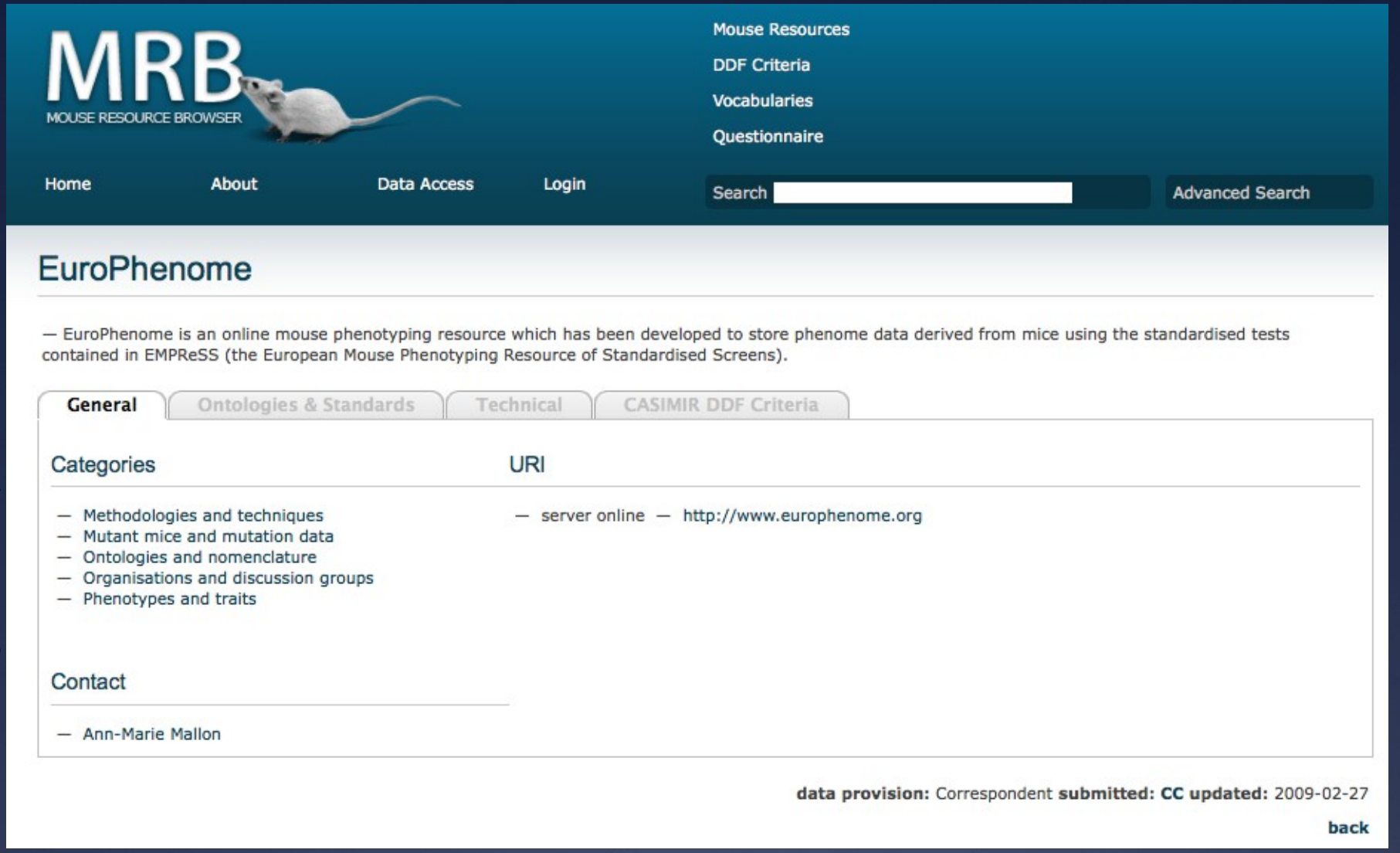
- Original meeting on 7-8 March 2008 in Cambridge discussed the MRB and the basis of the DDF
- Aim was a simple “traffic light” description of databases (NOT a quality assessment)
- CASIMIR-ENFIN workshop 30-31 October 2008, Berlin
- Presented on CASIMIR web site
- Smedley et al (2010) *Database* **2010:** baq014

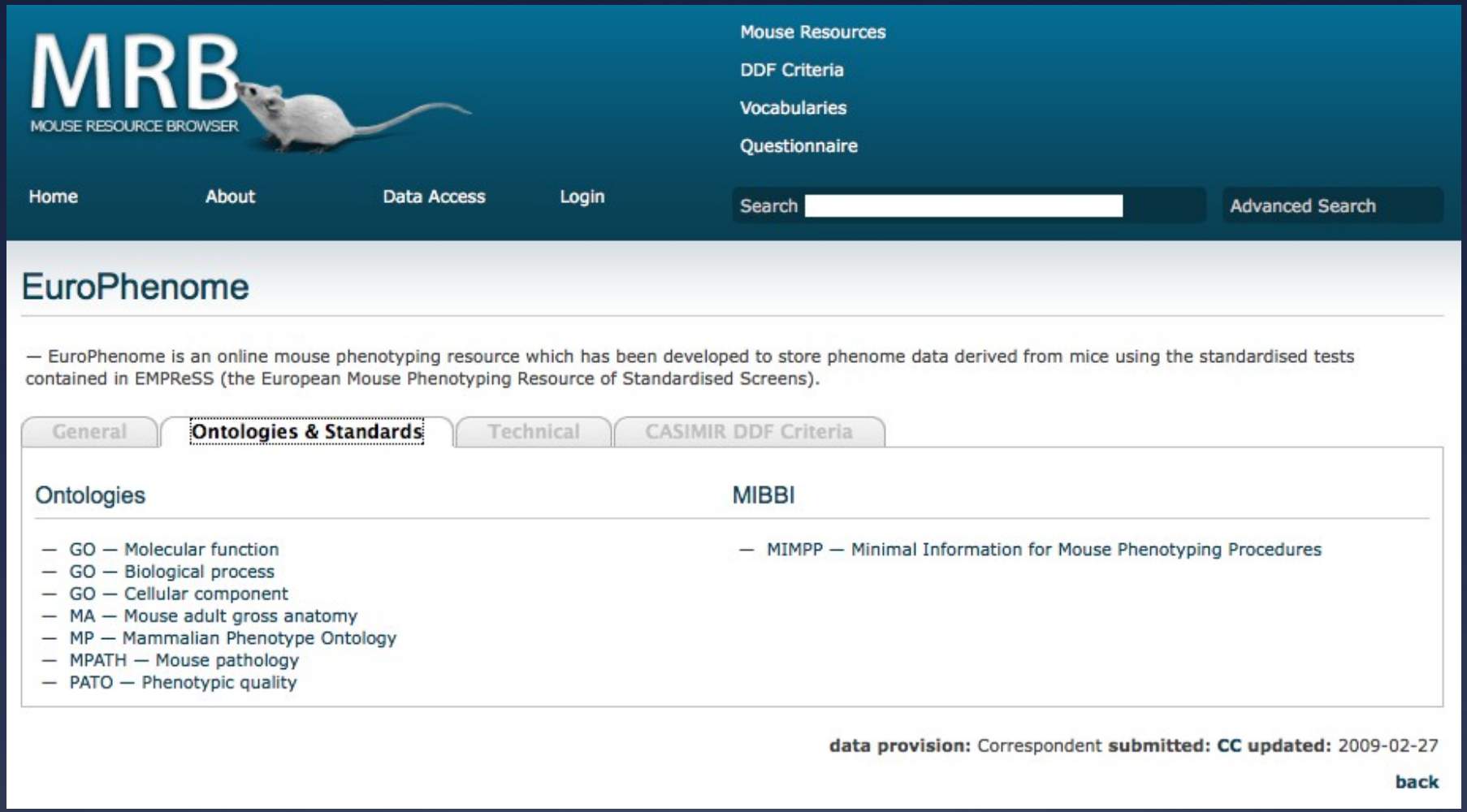
Mouse Resource Browser

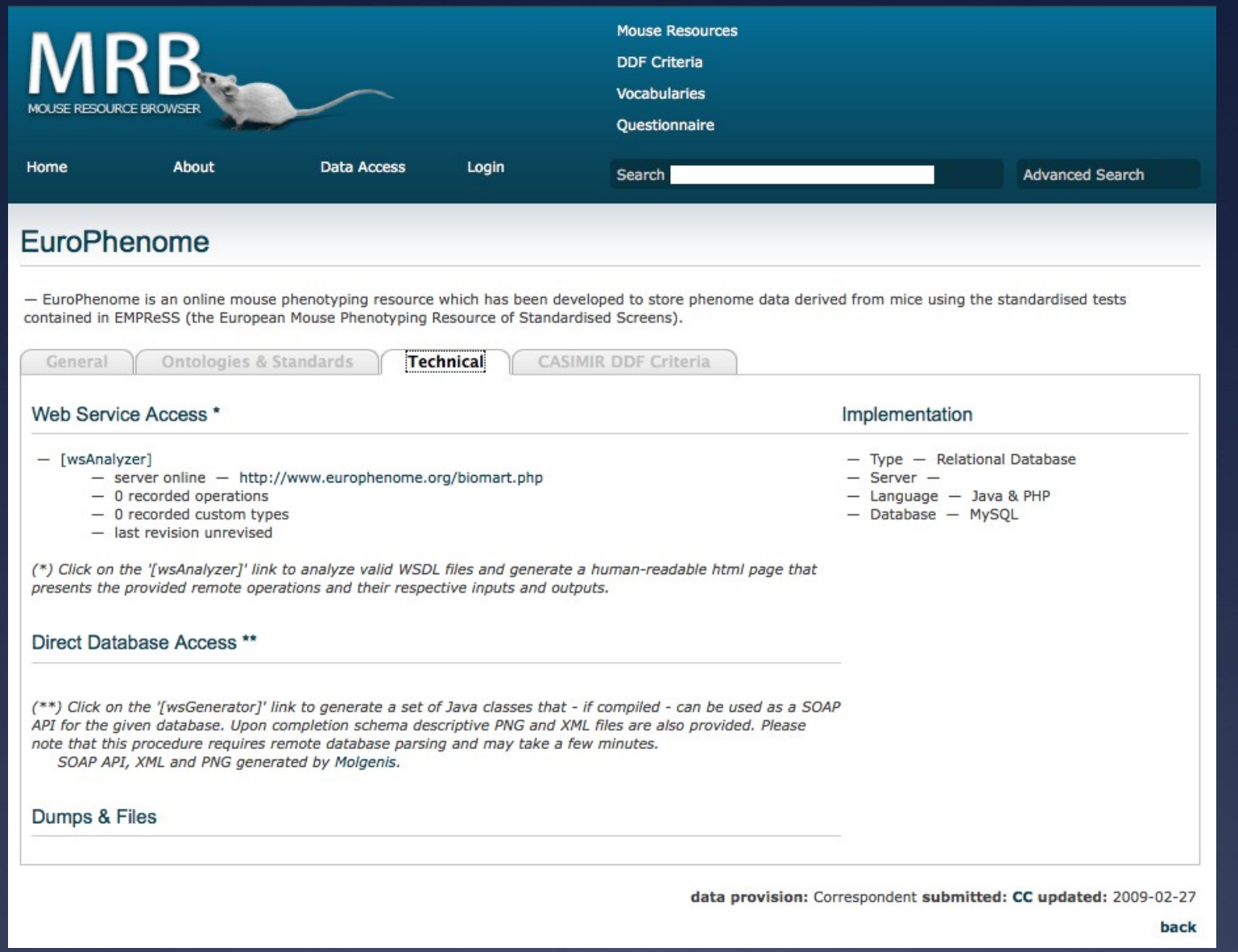


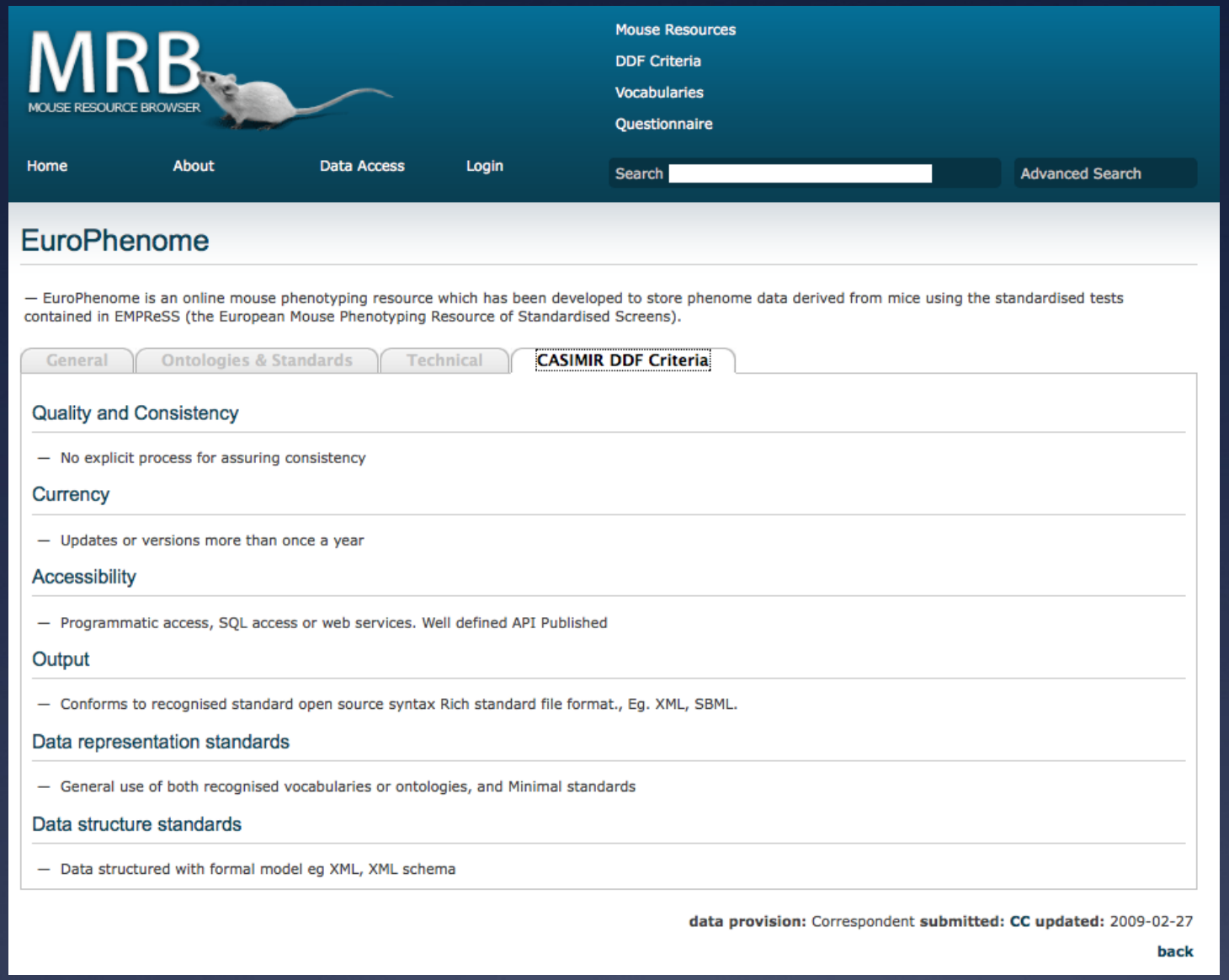
Christina Chandras
Michael Zouberakis
Vassilis Aidinis



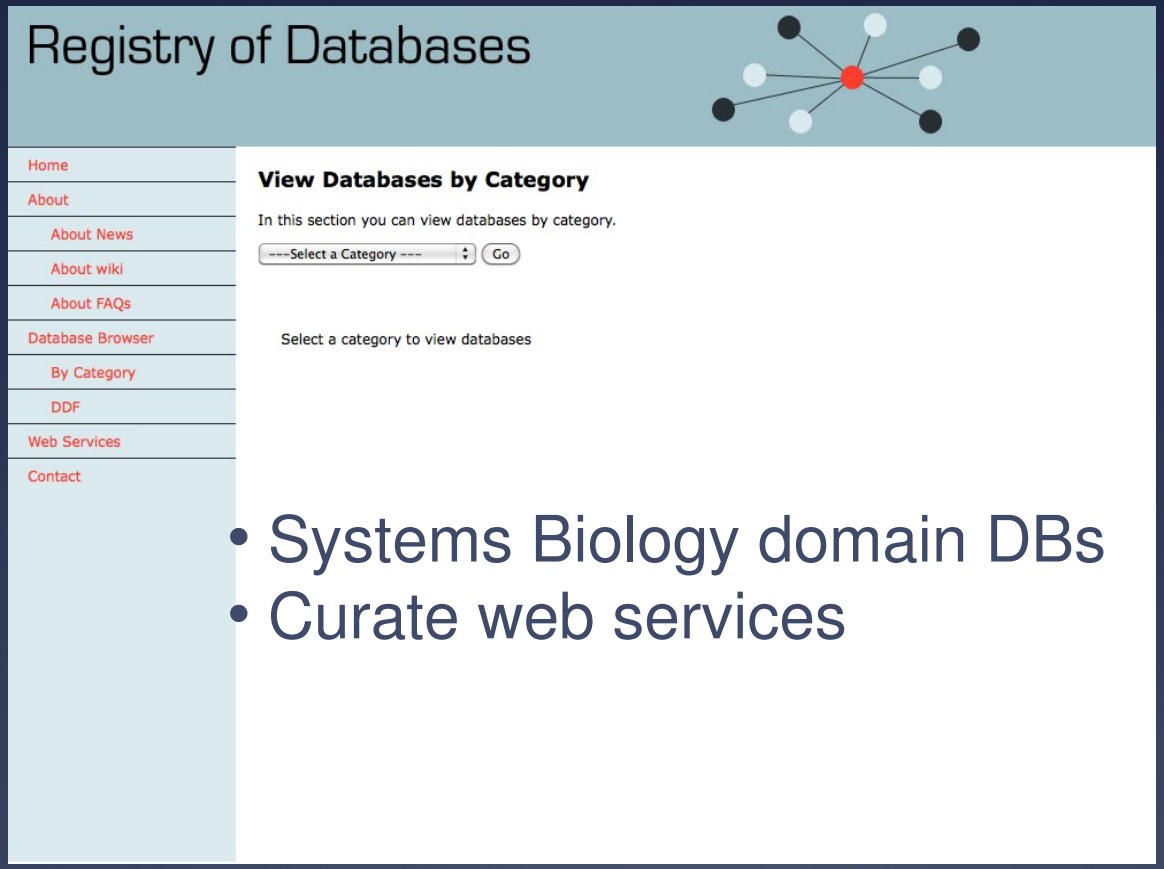








ENFIN Registry



- Systems Biology domain DBs
- Curate web services

Chrysanthi Ainali

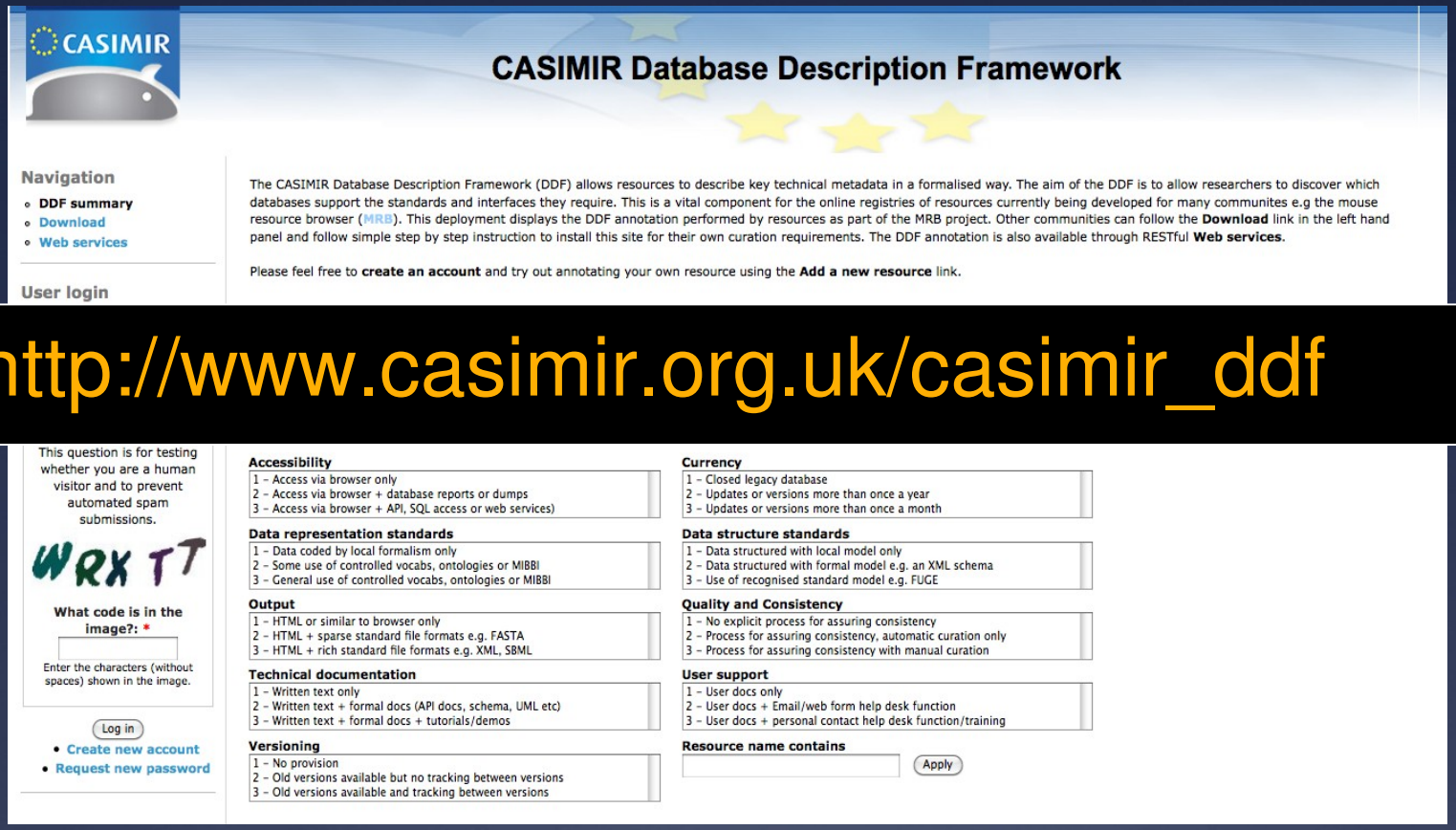
DDF

Database, Vol. 2010, Article ID baq014, doi:10.1093/database/baq014

Table 1. The CASIMIR Database Description Framework (DDF)

Category	Level 1	Level 2	Level 3
Quality and Consistency	No explicit process for assuring consistency	Process for assuring consistency, automatic curation only	Process for assuring consistency with manual curation
Currency	Closed legacy database or last update more than a year ago	Updates or versions more than once a year	Updates or versions more than once a month
Accessibility	Access via browser only	Access via browser and database reports or database dumps	Access via browser and programmatic access (well defined API, SQL access or web services)
Output formats	HTML or similar to browser only	HTML or similar to browser and sparse standard file formats, e.g. FASTA	HTML or similar to browser and rich standard file formats, e.g. XML, SBML (Systems Biology Markup Language)
Technical documentation	Written text only	Written text and formal structured description, e.g. automatically generated API docs (JavaDoc), DDL (Data Description Language), DTD (Document Type Definition), UML (Unified Modelling Language), etc.	Written text and formal structured description and tutorials or demonstrations on how to use them
Data representation standards	Data coded by local formalism only	Some data coded by a recognised controlled vocabulary, ontology or use of minimal information standards (MIBBI)	General use of both recognised vocabularies or ontologies, and minimal information standards (MIBBI)
Data structure standards	Data structured with local model only	Data structured with formal model, e.g. an XML schema	Use of recognised standard model, e.g. FUGE
User support	User documentation only	User documentation and Email/web form help desk function	User documentation as well as a personal contact help desk function/training
Versioning	No provision	Previous version of database available but no tracking of entities between versions	Previous version of database available and tracking of entities between versions

Database Description Framework



Damian Smedley

DATABASE
The Journal of Biological Databases and Curation

Database, Vol. 2010, Artide ID baq014, doi:10.1093/database/baq014

Original article

Finding and sharing: new approaches to registries of databases and services for the biomedical sciences

Damian Smedley[1,*], Paul Schofield[2], Chao-Kung Chen[1], Vassilis Aidinis[3], Chrysanthi Ainali[4], Jonathan Bard[5], Rudi Balling[6], Ewan Birney[1], Andrew Blake[7], Erik Bongcam-Rudloff[8], Anthony J. Brookes[9], Gianni Cesareni[10], Christina Chandras[3], Janan Eppig[11], Paul Flicek[1], Georgios Gkoutos[12], Simon Greenaway[7], Michael Gruenberger[2], Jean-Karim Hériché[13], Andrew Lyall[1], Ann-Marie Mallon[7], Dawn Muddyman[2], Florian Reisinger[1], Martin Ringwald[11], Nadia Rosenthal[14], Klaus Schughart[15], Morris Swertz[16], Gudmundur A. Thorisson[9], Michael Zouberakis[3] and John M. Hancock[7]

Thanks To

Nature Precedings : doi:10.1038/npre.2011.6158.1 : Posted 26 Jul 2011

CASIMIR

- Paul Schofield
- Damian Smedley
- Christina Chandras
- Michael Zouberakis
- Vassilis Aidinis
- Dawn Muddyman

ENFIN

- Manuel Corpas
- Chrysanthi Ainali
- Pascal Kahlem
- Ewan Birney

All participants in CASIMIR and CASIMIR-ENFIN workshops