CASIMIR: Syntactic and Semantic Interoperability in Mouse Databases

John Hancock

MRC Harwell and

The CASIMIR Consortium

MRC | Harwell

Mammalian Genetics Unit

Why Interoperability?

- Facilitates data sharing
- Allows for design of more sophisticated search and analysis portals
- Facilitates comparative analysis across species

CASIMIR

MRC Harwell
Mammalian Genetics Unit

Project

- FP6 Coordination Action; €1.3M for three years
- Follow-on from the PRIME initiative
- Started February 2007
- Brings together ten partners from four different European countries representing the interests and experience of Universities, Research Institutes and Industry
- Involves the whole scientific community in a consultation exercise on informatics needs for mouse functional genomics
- Underpins the European Research Area in mouse functional genomics

CASIMIR

MRC Harwell

Mammalian Genetics Unit

Partners

- **University of Cambridge**- Coordinator
- **Medical Research Council**
  - Mammalian Genetics Unit, Harwell, UK
  - Human Genetics Unit, Edinburgh, UK
- **EMBL**
  - EBI, Hinxton, UK
  - EMBL-Monterotondo Outstation, Italy
- **Helmholtz-zentrum fuer infektionsforschung Braunschweig**,Germany
- **Helmholtz- zentrum Munich**
- **Biomedical Research Centre**, **Alexander Fleming**, Athens, Greece
- **Consiglio Nazionale delle Richerche**, Rome, Italy

CASIMIR

MRC | Harwell
Mammalian Genetics Unit

CASIMIR Work Packages

- WP1 Management
- WP2 Dissemination
- WP3 International Advisory Group
- WP4 Data representation and semantics
- WP5 Database compatibility and interoperability
- WP6 Data acquisition, curation and ownership
- WP7 Integration of Biological Collections and physical resources into the data network
- WP8 User Interactions with data and databases

CASIMIR

MRC Harwell

Mammalian Genetics Unit

Requirements for Syntactic and Semantic Operability

- Speak the same language
  - Controlled vocabularies and ontologies
  - Standard XML(s) for data transfer
- Databases “integration-ready”
  - Web services/other direct access
  - Or BioMart integration

CASIMIR

MRC Harwell
Mammalian Genetics Unit

CASIMIR Questionnaire

28 databases targeted with an on-line questionnaire concerning:

- RDBMS, object, other
- Web services
- Ontologies
- Minimum standards
- Future needs

MRC Harwell

Mammalian Genetics Unit

Conclusions of Questionnaire

- Most DBs now using **RDBMS** - mostly MySQL or PostgreSQL
- **Web services** are a popular option amongst the “EU” databases although less than half have implemented them so far. The other DBs show less penetration of web services
- **Ontologies** are in wide use but relatively poorly used in “EU” databases. Supplementation with CVs more common in non-”EU” databases
    - Many, but not all, DBs not using ontologies will or are considering using them. Many DBs currently using ontologies will also extend their ontology use
    - Although OBO ontologies are used by most ontology users these only make up a slim majority
    - Most DBs see a need for new ontologies (often undefined)
- **Minimum standards** are currently in use by only c. 25% of DBs - the most commonly used is MAGE/MIAME

CASIMIR

MRC Harwell
Mammalian Genetics Unit

Syntactic Integration

- CASIMIR currently developing use cases using a mixture of integration frameworks:
  - BioMart
  - MOLGENIS
  - Taverna (also: EnSUITE)
- Paper in press in Briefings in Bioinformatics: Smedley et al

CASIMIR

MRC Harwell
Mammalian Genetics Unit

Semantic Interoperability

- Many shared ontologies in use, e.g. Gene Ontology
- A particular area of interest in the mouse community is phenotype ontologies
- We are supporting discussions on relating mouse phenotype to human disease

CASIMIR

MRC Harwell
Mammalian Genetics Unit

CRITERION	1 star	2 stars	3 stars
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 legacydatabase	Updates or versions more than once a year	Updates or versions more than once a month
Accessibility	Access via browser	Database reports or database dumps	Programmatic access. SQL access or web services. Well defined API Published
Output	Conforms to recognised standard open source syntax; html or similar to browser	Conforms to recognised standard open source syntax Sparse standard file format. Eg. FASTA	Conforms to recognised standard open source syntax Rich standard file format., Eg. XML, SBML.
Technical documentation	Written text	Formal structured description, eg automatically generated API eg JavaDoc, schema, UML etc	Tutorials, demonstrations, plus ** criteria
Data representation standards	Data coded by local formalism only	Some Data coded by recognised controlled vocabulary or ontology or use of MIBBI	General use of both recognised vocabularies or ontologies, and Minimal standards
Data structure standards	Data structured with local model	Data structured with formal model eg XML, XML schema	Use of recognised standard model, eg FUGE
User support	User documentation	Email/web form help desk function	Personal contact help desk function/training

CASIMIR's Database of Databases (Mouse Resource Browser)

- What databases are available?
- What kinds of data do they make available
- Do they use commonly accepted standards (e.g. ontologies, MIs)
- Do they make data available via web services or other programmable means?

CASIMIR

MRC Harwell
Mammalian Genetics Unit

Thanks to…

Paul Schofield (Coordinator)

Vassilis Aidinis, Christina Chandras, Michael Zouberakis (WP7 and organising this workshop)

Damian Smedley (WP5)
Nadia Rosenthal (WP6)
Klaus Schughart (WP8)

The CASIMIR Consortium

…and everyone else who has contributed to our meetings so far

MRC Harwell
Mammalian Genetics Unit

CASIMIR Annual meeting 2008

Nobel Forum, Karolinska Institute, Stockholm December 1-3 2008

CASIMIR

MRC Harwell

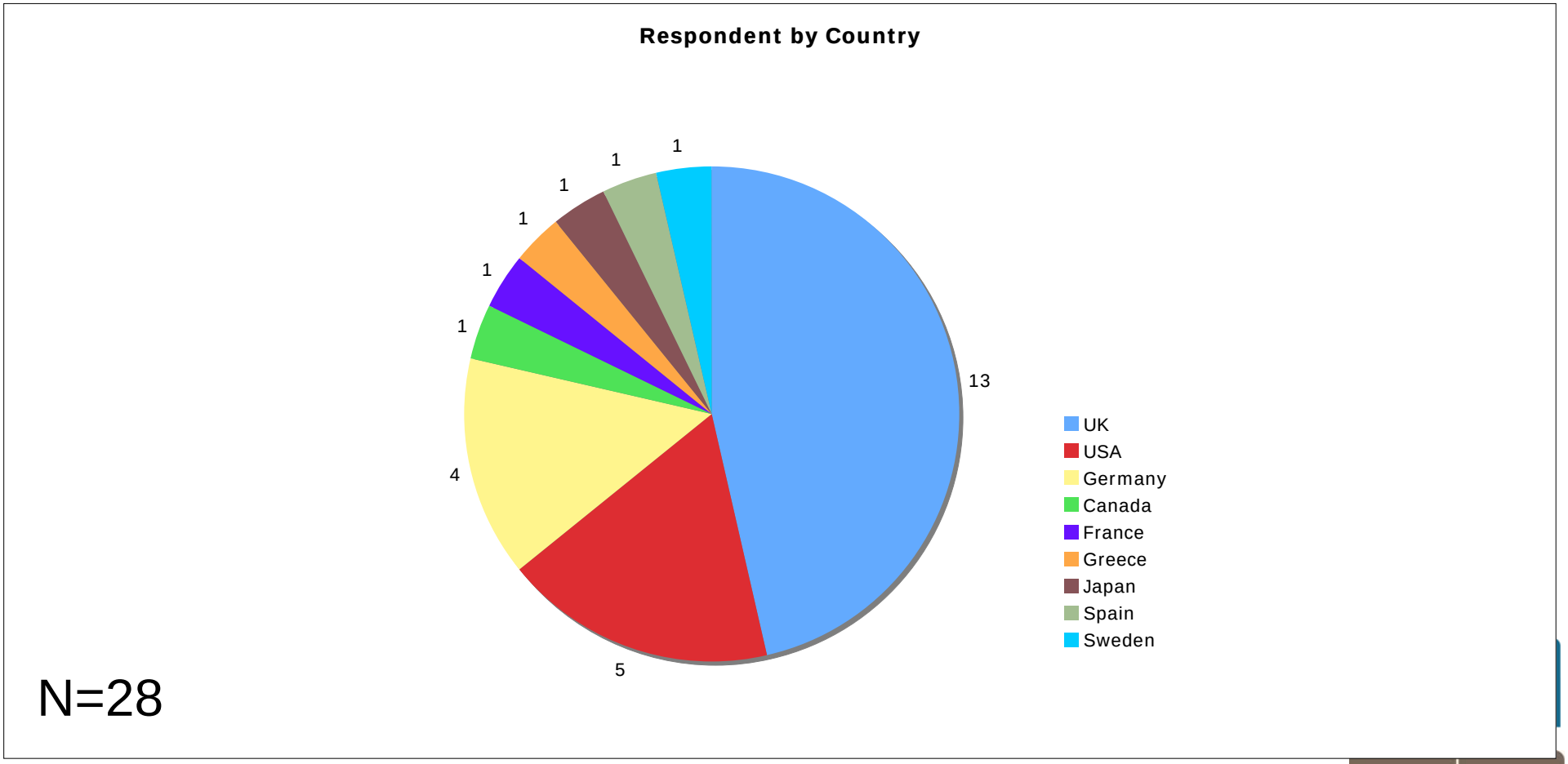
Mammalian Genetics Unit

CASIMIR Questionnaire Results

MRC | Harwell

Mammalian Genetics Unit

Overall Reponses

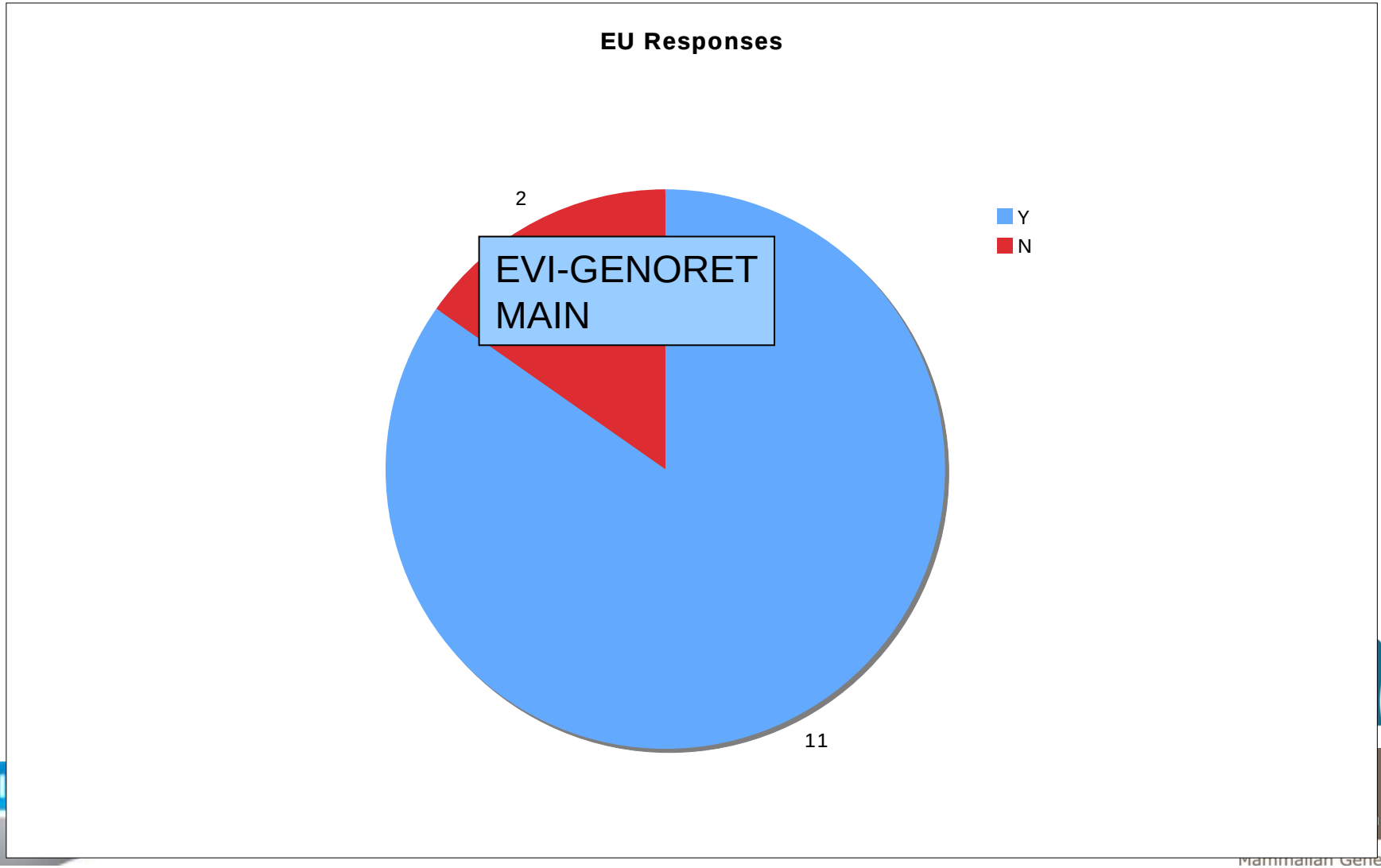


N=28

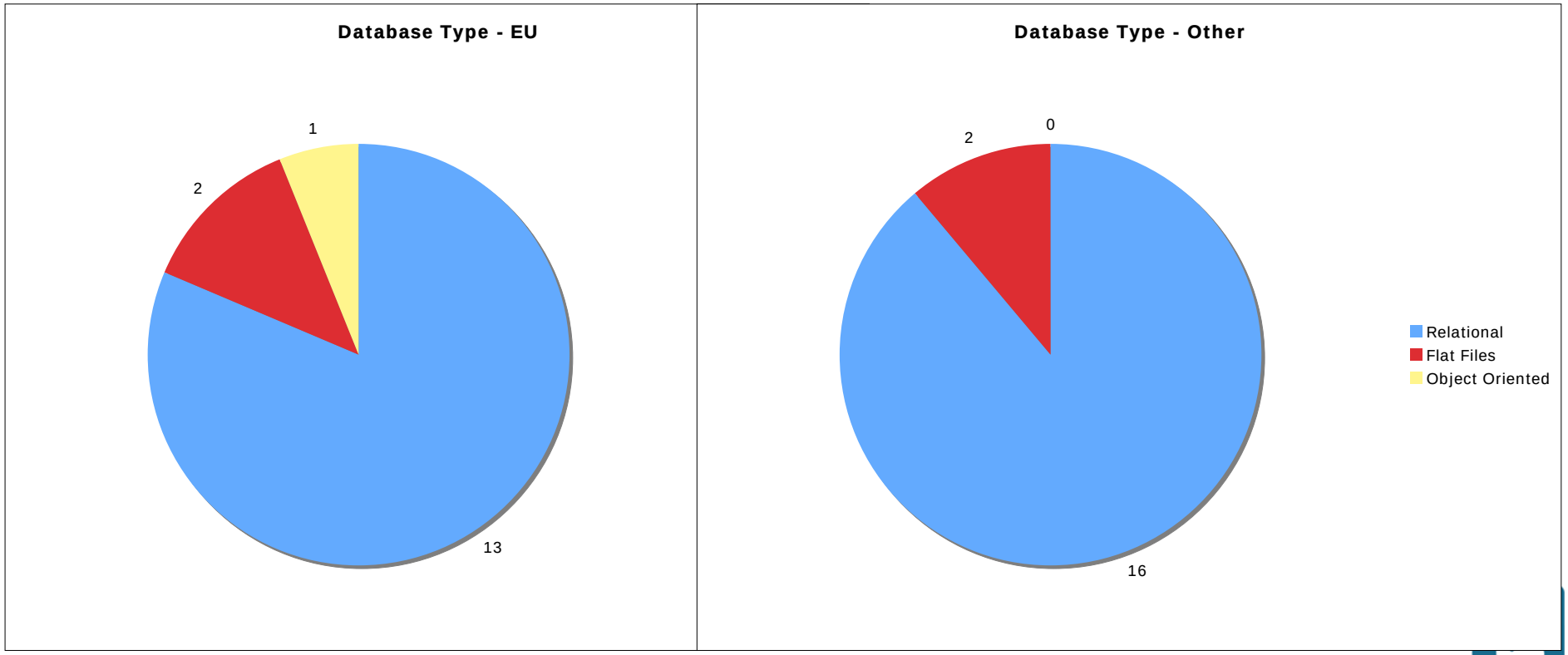
CASIMIR

MRC Harwell
Mammalian Genetics Unit

EU Project Response



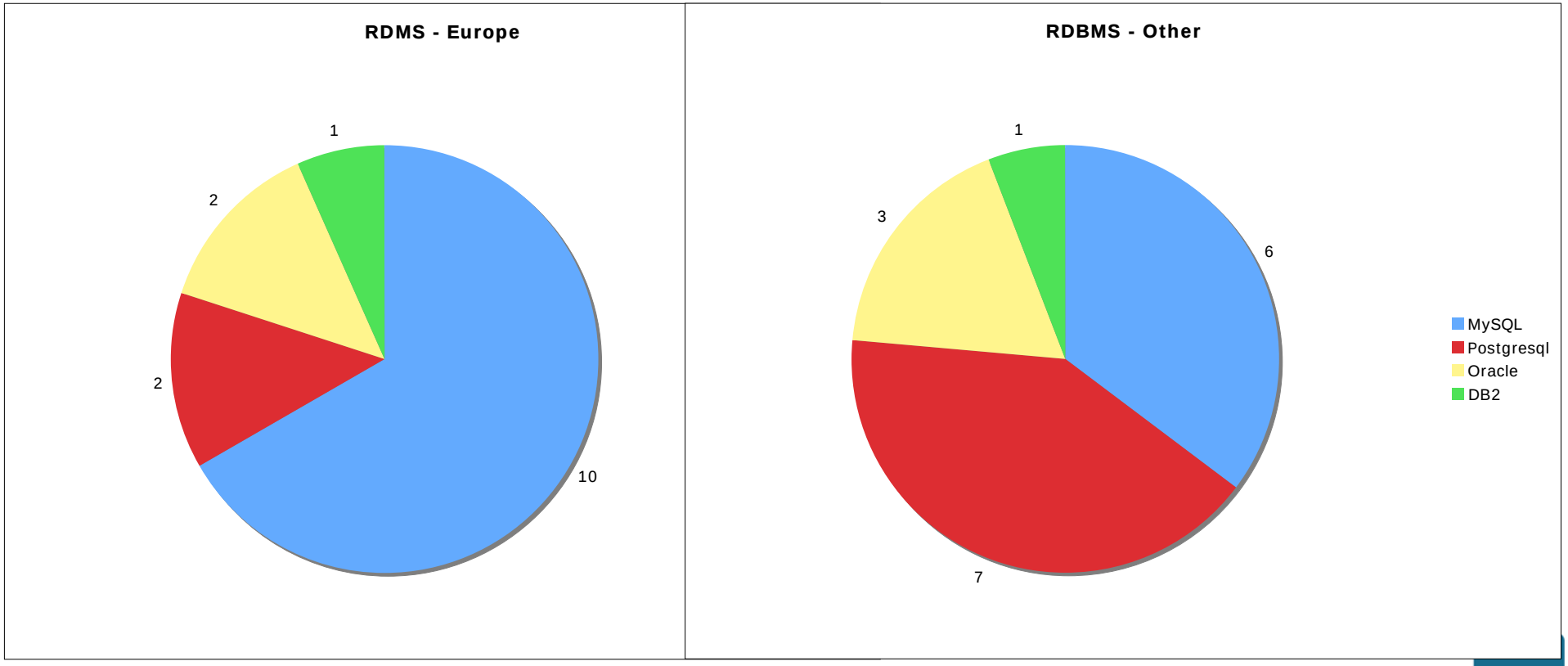
Are you using a relational database, object database or flat files?



CASIMIR

MRC Harwell
Mammalian Genetics Unit

If relational, what is your chosen Relational Database Management System (such as MySQL, Postgresql, Oracle)?

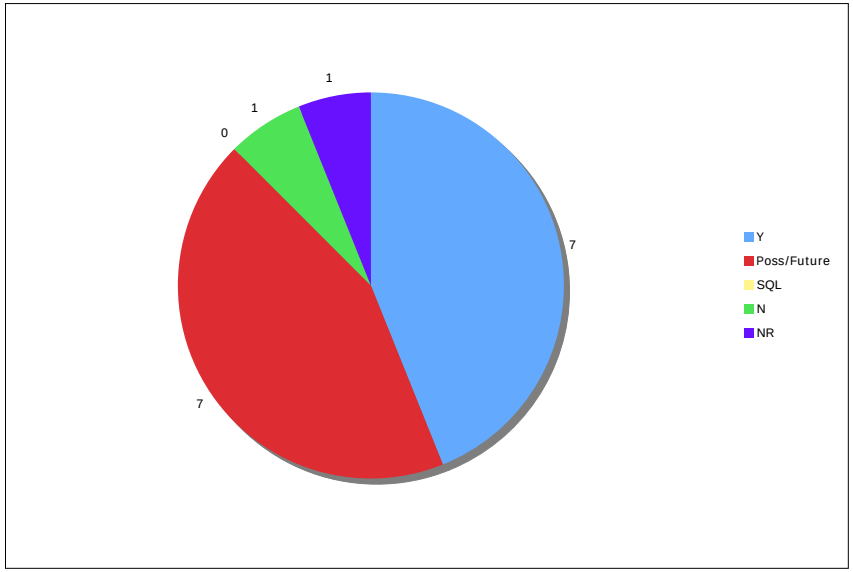


MySQL predominates in EU DBs, whereas PostgreSQL is more popular with non-EU users

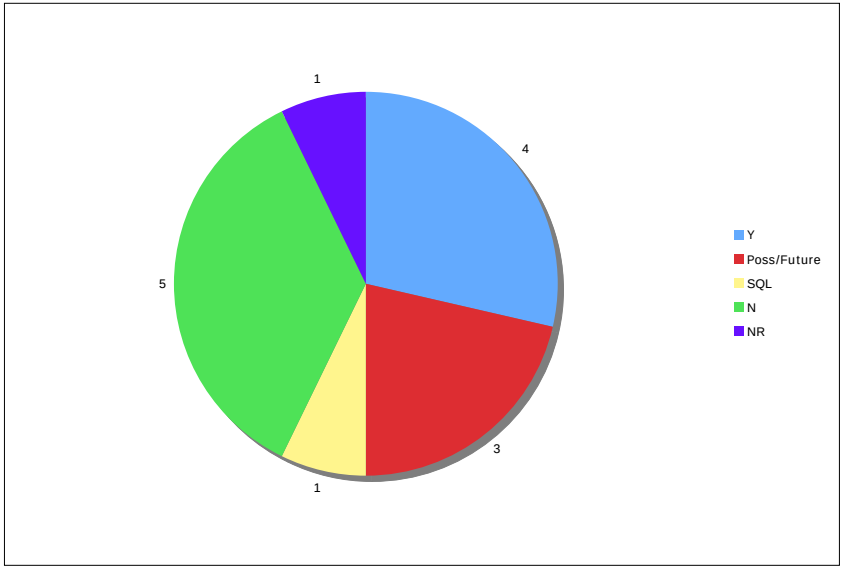
MRC Harwell
Mammalian Genetics Unit

Supported/Installed Web Services (if yes please name them)? Do you plan to install or develop web services in the near future?

EU



Other



Web services more “popular” amongst “EU” databases

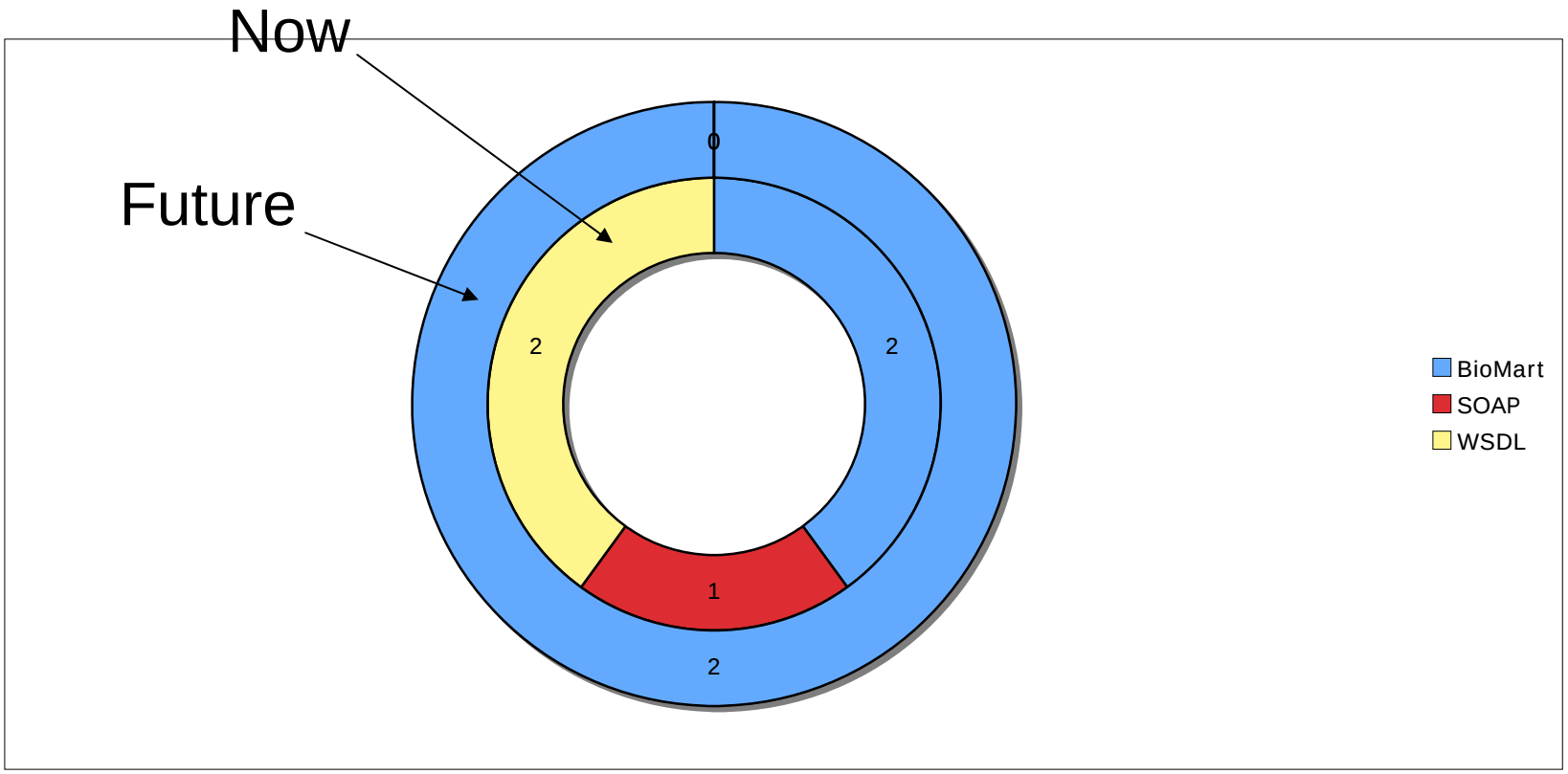
Do these provide data or results of analyses?

CASIMIR

MRC Harwell

Mammalian Genetics Unit

What kind of web service

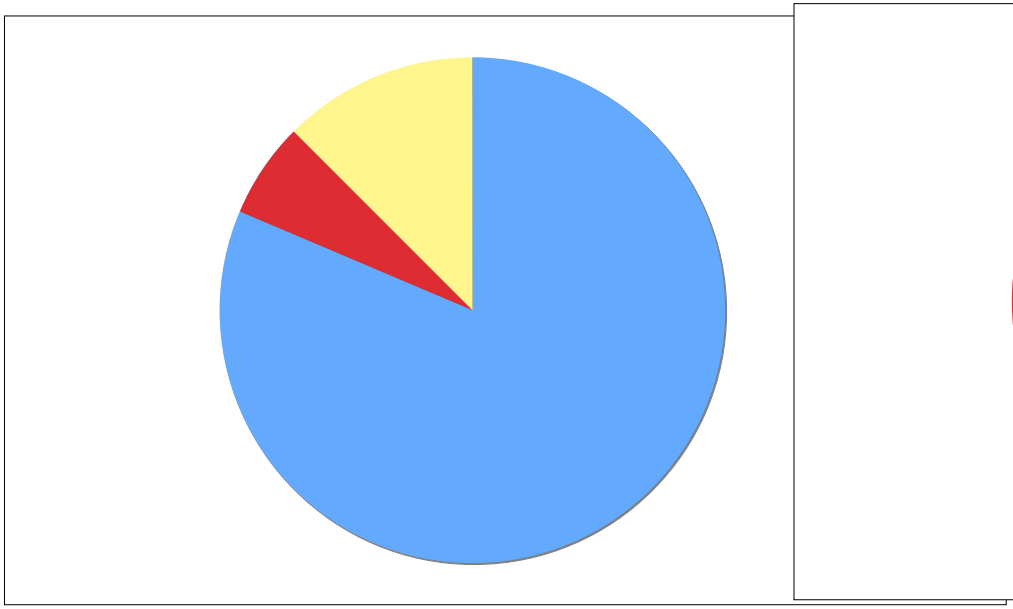


CASIMIR

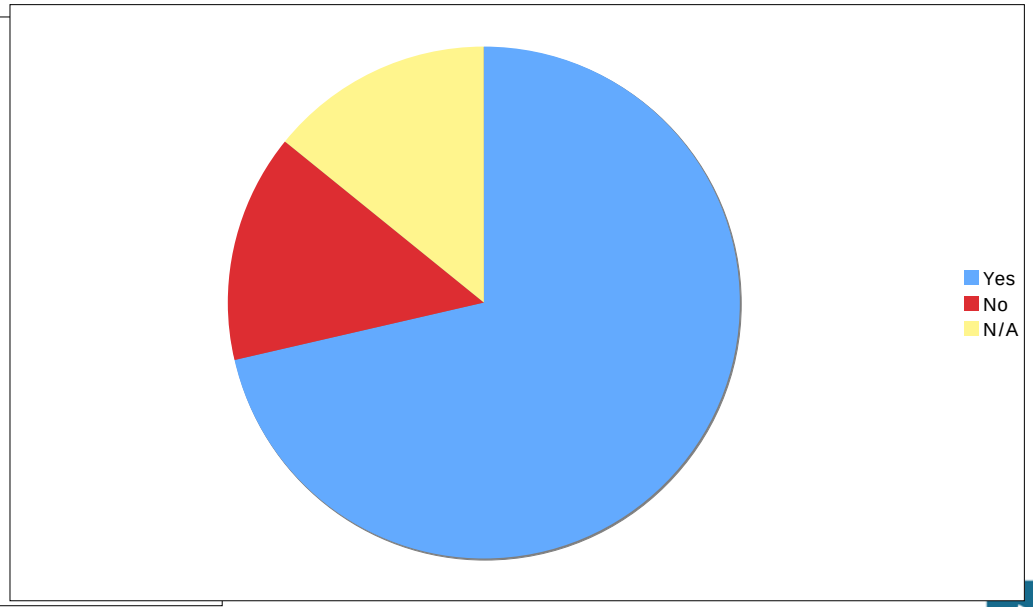
MRC Harwell
Mammalian Genetics Unit

Are you willing to provide a entity relationship diagram and would you be willing to provide it under an open source license?

EU



Other



Largely, yes

CASIMIR

MRC Harwell

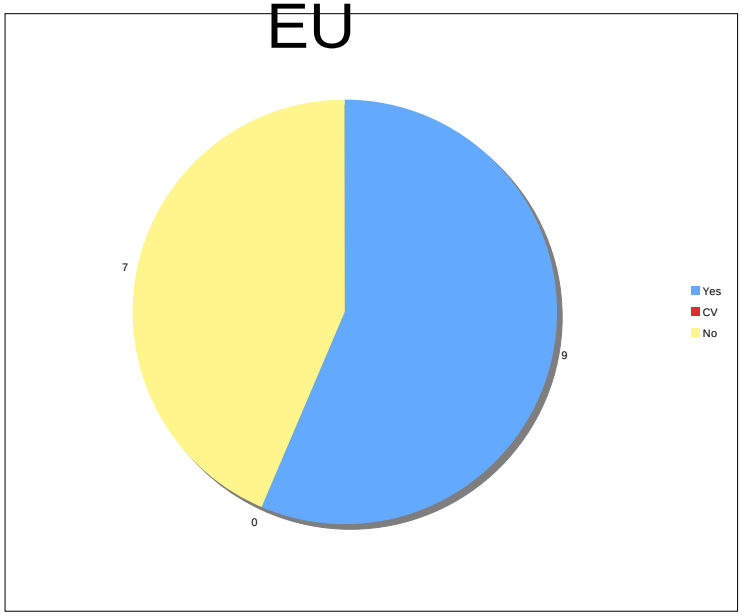
Mammalian Genetics Unit

Please list the sorts of data entities you store (e.g. protein sequence data, mouse strain information etc...)

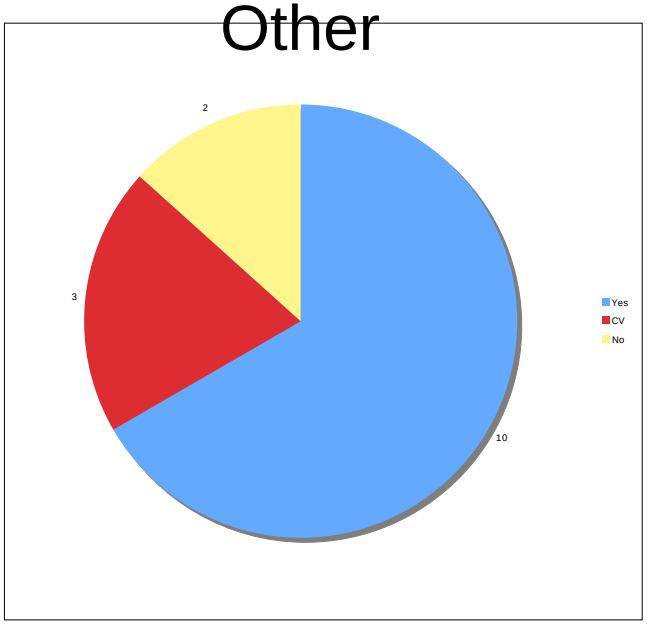
Phenotype	SOPs
Genotype	Resequencing data; SNPs; CNVs; Mutation types
Strain information	Strain availability; mouse requests; cryopreservation; husbandry data
Gene information	Sequence; allele information; gene models; supporting evidence; predicted regulatory features; sequence annotation
Gene product information	Transcript; protein sequence; protein annotation
Targeted mutation information	Knockout design; DNA constructs; allele requests; microinjection pipelines; ES cells & QC results
Gene expression data	Gene regulation information; qRTPCR results; gene expression platform metadata; coexpression cluster information
Spatial gene expression data	In situ probes; antisera
Images	
Ontology terms & definitions	
Pathology data	
Researcher information	
Cell lines	
Systems biology	Models; pathways
Publications	

Are you currently using or do you intend to use any ontologies or controlled vocabularies to describe your data (if yes please name them)?

EU



Other

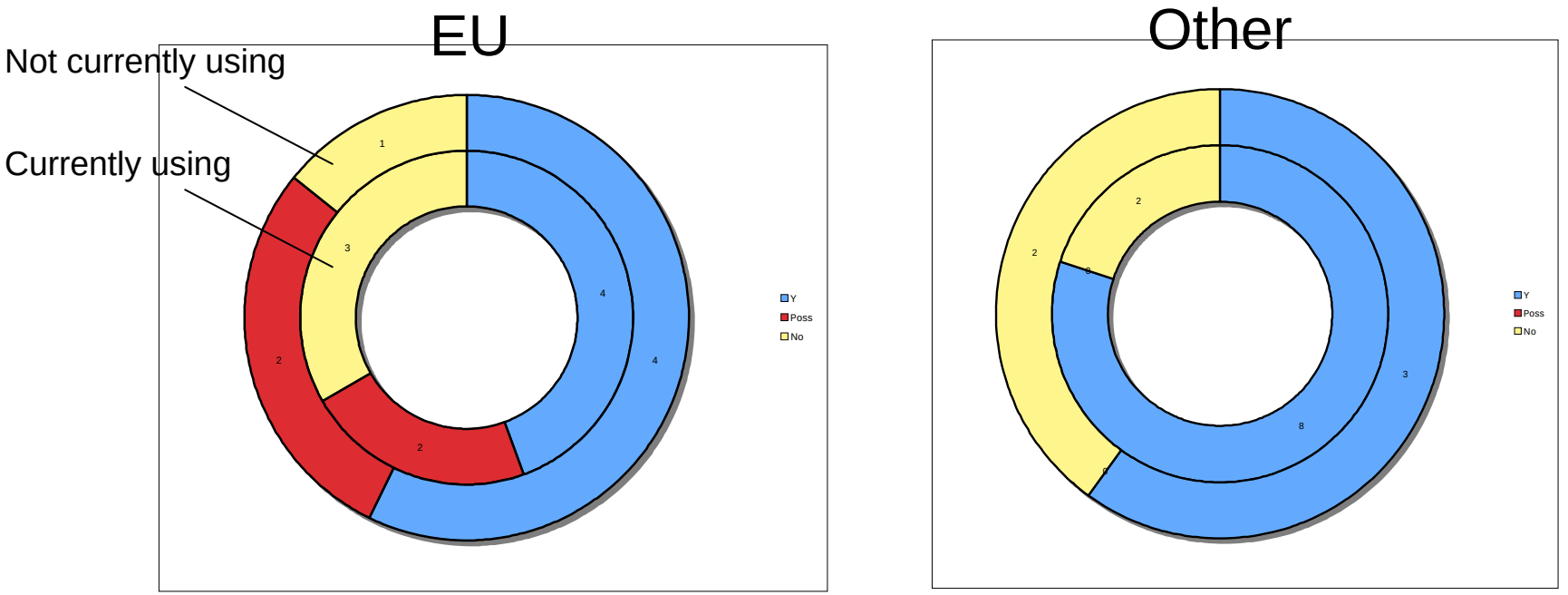


Greater use of CVs to supplement absence of ontologies in “other” group

CASIMIR

MRC Harwell
Mammalian Genetics Unit

Do you plan to expand your use of ontologies in future?

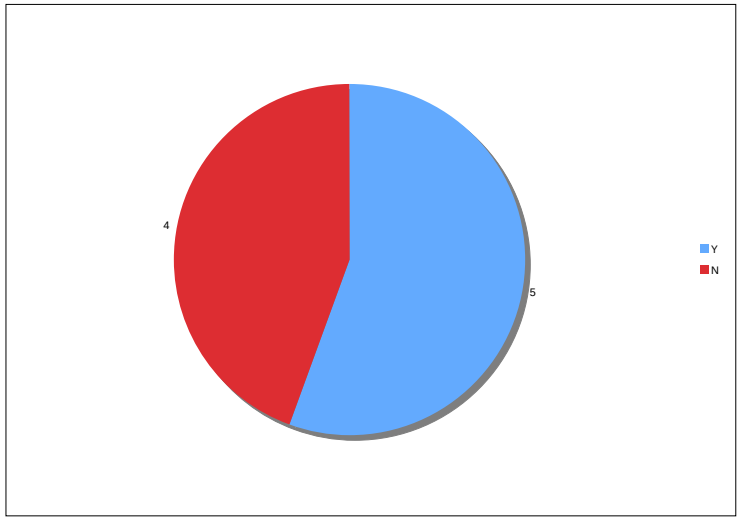


- Most DBs using ontologies will or are considering extending their use
- Most DBs not using them will adopt them or are considering doing so

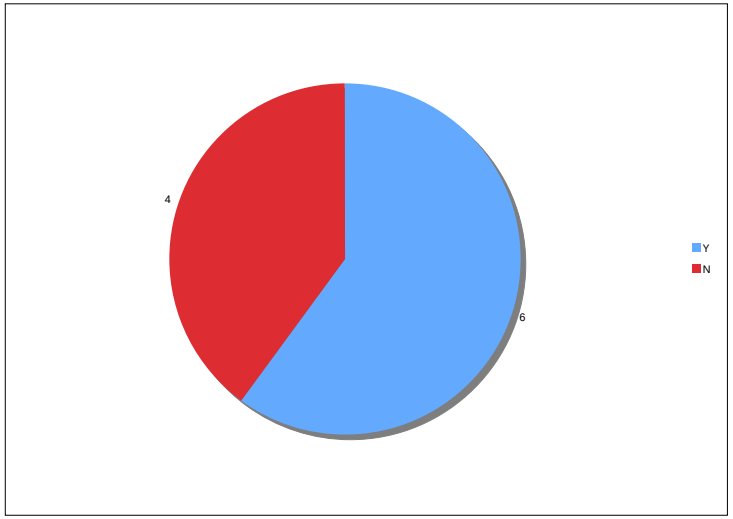
MRC Harwell
Mammalian Genetics Unit

Do you use OBO ontologies?

EU



Other



A majority, but not all, use OBO ontologies

Clearly, there is a strong argument for standard ontologies

CASIMIR

MRC Harwell

Mammalian Genetics Unit

Named OBO Ontologies in Use

ChEBI (2+1)	
CL (Cell type) (1+1)	
EMAP (2)	
GO (4+3)	
Human Anatomy (?)	
MA (Mouse Anatomy) (2+2)	
MGED (MO)	
MP (3+1)	Custom CV for phenotypes (MausDB)
MPATH	
NCBI Taxonomy (taxon)	
OBI (2)	
PATO (1+1)	
SO (+2)	
ZFIN	

MRC | Harwell

Mammalian Genetics Unit

Named non-OBO Ontologies in Use

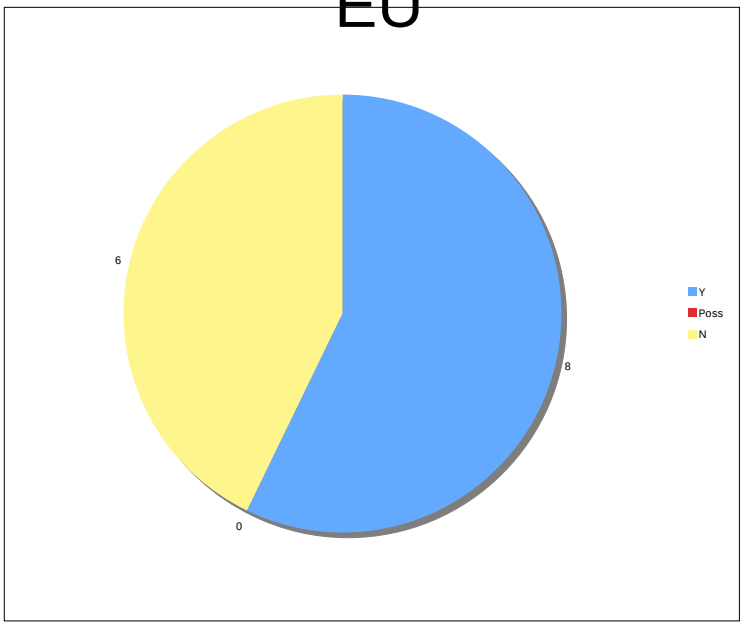
BIRNLex
GSC Mouse Phenotyping Experiments (GSCMPE)
Mouse diagnoses (generated by the pathologists of the Mouse Models of Human Cancers Consortium)
NCI EVS ontology
Organ, tissue and cell structure
Rat anatomy (deducted from RENI)
Rat diagnoses of proliferative diseases (from RENI)
Staining methods
V gene sequence regions (CV)

MRC Harwell

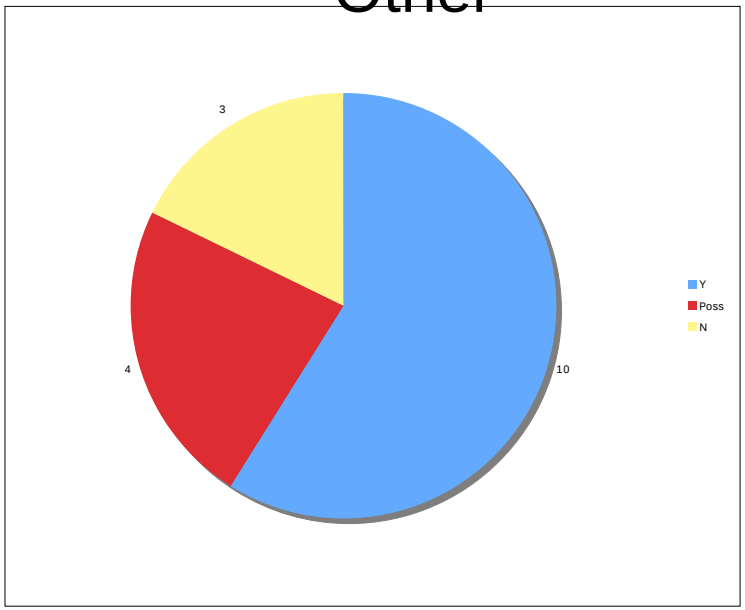
Mammalian Genetics Unit

Do you perceive the need for additional ontologies to serve your domain of knowledge?

EU



Other



A majority see the need for new ontologies or think this may be needed

CASIMIR

MRC Harwell
Mammalian Genetics Unit

Named Ontologies Needed

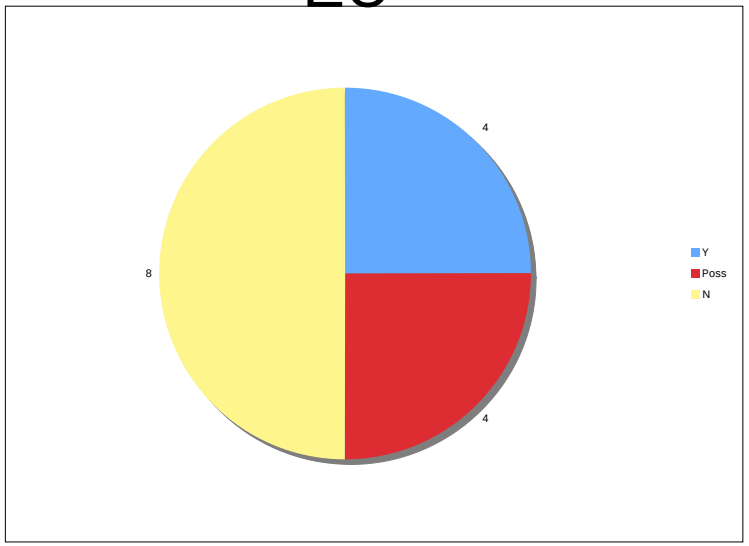
- Modified RNA/DNA molecules
  - "An ontology for RNA and DNA molecules with various characteristics, e.g. capped, methylated mRNA, nascent mRNA, DNA with 3` overhanging end, etc."
- Phenotype ontology (!)
- Experimental setups/SOPs
- General anatomy (viz. CARO)
- Gene products (?)

CASIMIR

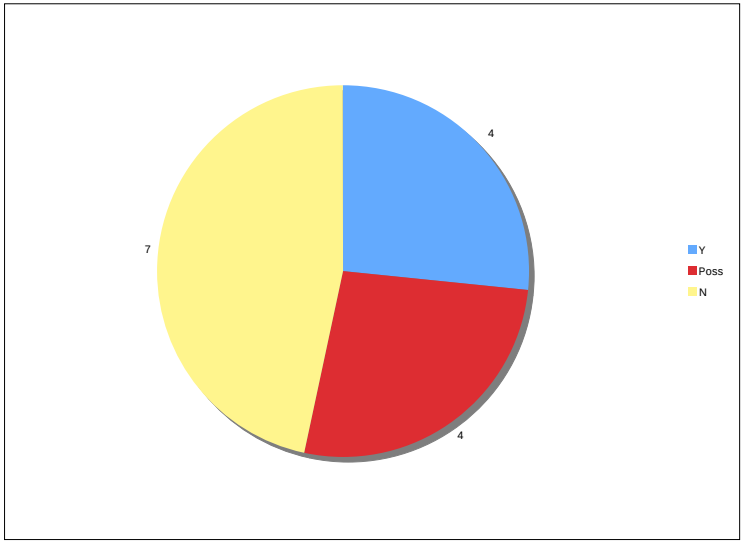
MRC Harwell
Mammalian Genetics Unit

Do you make use of Minimum Information standards (such as MIAME for microarray experiments) to describe any data? If so, which ones? If you do not make use of these standards, are you likely to do so in future?

EU



Other



Use of minimum standards is limited although a number of DBs are considering adopting them or starting to do so

CASIMIR

MRC Harwell
Mammalian Genetics Unit

Minimum Standards In Use

MIAME/MAGE

MISFISHIE

(MIMPP - Mouse phenotyping experiments)

CASIMIR

MRC Harwell

Mammalian Genetics Unit

What do you perceive as the main limiting factor in data representation/interoperability etc. in European bioinformatics databases?

•Determining equivalence. e.g. what Uniprot protein does a certain Ensembl gene really correspond to? How does one resolve ambiguities?
•Use of multiple identifiers. For databases addressing similar contents use of different semantic and schemas/data models.
•Lack of standards for many kinds of experimental data
•Databases built to perform a specific function / solve a specific problem, with little regard to interoperability. Lack of inter-group communication.
•The absence of an standard to represent the data. It should exist a minimal information schema compatible across all databases and then, each database, could extend that minimal schema with her owns needs
•Time (and money for someone to do this!)
•Inconsistensies in gene annotation
•We need a clear federating architecture. One-stop-shops will not cut it: the best tactic is to leave the data searching / presentation to each expert group (whether that is gene-model reconciliation, es cell production, phenotyping etc), and have a well-defined federating protocol between each data store. The strength of this approach is that you can run true cross-store searches, which greatly increases the utility of each store.
•Semantic integration issues. International compatibility e.g. with caBIG, NCBI.
•Lack of clear vision for exactly what is needed and who will provide it. Spreading the workload is inclusive, but it can lead to a less well formed product.
•1) To get the DAS Server and running again 2) Access to phenotype databases via webservices

MRC Harwell
Mammalian Genetics Unit

What do you perceive as the main limiting factor in data representation/interoperability etc. in European bioinformatics databases?

•We don`t need more and more distributed databases describing the same things again and again just with different IDs. We need integration platforms that collect and non-redundantly present data from distributed databases. But: the integration platform must not copy data from other databases and thereby create a new database. I should be more a search engine like NCBI Entrez cross database search (http://www.ncbi.nlm.nih.gov/gquery/gquery.fcgi) Limiting factor to achieve this might be common data structures and lack of web services.
•Incompatible approaches to data integration and representation (e.g. DAS servers vs. flat files; UML; various incompatible flavors of web service, etc, etc); the `NIH` (`Not Invented Here`) syndrome. Identifier mapping is also a perennial issue.
•Interoperability is primarily a matter of agreement on standards, incl. ontologies; there remains a need for agreed, widely-accepted ontologies;
•`Database of databases` would help the user to find his / her way - web service availabilty should get a common standard - standard for descriptions of web services - standard ontology for biological web services (as in BioMoby)
•Lack of common standards * lack of use of webservices * data privacy / concurency
•Insufficient penetration of database expertise even in large centers (particularly on the continent; England is the exception).
•There are too many ontologies and little agreement between them with respect to similar concepts

CASIMIR

MRC Harwell
Mammalian Genetics Unit

Do you have any comments/thoughts on standards for data representation that need to be developed or that you might like discussed in CASIMIR?

•Standardised identifiers/mappings
•standards for experimental data
•Standards only work if they are adopted by everyone.
•We need to find what information needs to be shared. Then all of our databases must implement this very simple format in order to be interoperable among databases. Then, each database could extend that simple format to be able to fulfill his aims. Just like object oriented programming. We need the base class and then extend it to make more complex data.
•For in situ expression data, MISFISHIE standards should be discussed/employed
•Semantic integration issues, collaborative ontology development.
•For the virtual mouse Webservices for phenotype databases, or adding `a virtual mouse` to the phenotype database to visualise the phenotypic data.
•Original source and evidence level should alway be displayed or at least made available on request (`click here to see evidence` -> `curated author statement`, `inferred from electronic annotation`,...)
To be further developed: visualisation To be developed: Reasoning support. An example: - user requests info about a gene - first level: display everything from distributed databases as usual - next level: cross check data. Is expression data consistant with promotor organisation data? Are co-citations from literature (gene-tissue/organ) consistant with expression data? What does this mean: not only databases need to be integrated and made be available via web services but also bioinformatics tools (literature mining tools, promotor analyis tools, ...)
•Extending anatomy ontologies into spatio-temporal aspects of geometric models;
•Recommendations of `preferred` ontologies ? - define standards fields to describe concepts (eg mouse line, gene, mutation ...) - propose a standard (such as SOAP, REST or XMLRPC) to deliver webservices based on those fields (use WSDL ?)

MRC Harwell
Mammalian Genetics Unit

Do you have any comments/thoughts on standards for data representation that need to be developed or that you might like discussed in CASIMIR?

•We are now developing a data format named as `Standardized Description of Operation Procedure (SDOP)` for the comparison of detailed parameters of phenotyping assays. We hope to collaborate with Europearn activities.
•Impact of RDF and semantic web technology.
•I would like to see universal adoption of a limited set of ontologies as well as see mechanisms put in place to make updating the ontologies feasible
•it would be nice to have a standard for represenation of mouse model (to begin with) and a general standard for representing animal models. It would also be nice to have a central location for vocabularies with easy programmatic access.
•Use the OBO ontologies, they will work with you to adapt as needed
•Iams?

CASIMIR

MRC Harwell
Mammalian Genetics Unit

Other (important) questions

Is your database providing external links to other on-line resources; possibly via URL/HTTP (if yes please name them)?

Can you provide a brief 'explanatory' description/schema of your data/data structure?

Do you have any comments/thoughts on standards for data representation that need to be developed or that you might like discussed in CASIMIR?

CASIMIR

MRC Harwell
Mammalian Genetics Unit

Conclusions

- Good response
- Most DBs now using RDBMS - mostly MySQL or PostgreSQL
- Web services are a popular option amongst the “EU” databases although less than half have implemented them so far. The other DBs show less penetration of web services
- Ontologies are in wide use but relatively poorly used in “EU” databases. Supplementation with CVs more common in non-”EU” databases
- Many, but not all, DBs not using ontologies will or are considering using them. Many DBs currently using ontologies will also extend their ontology use
- Although OBO ontologies are used by most ontology users these only make up a slim majority [Why?; What are perceived advantages of OBO?]
- Most DBs see a need for new ontologies (often undefined)
- Minimum standards are currently in use by only c. 25% of DBs - the most commonly used is MAGE/MIAME [Why?]

MRC Harwell

Mammalian Genetics Unit