BioMedBridges

Constructing data and service bridges in the life sciences

EUDAT 1st conference

Barcelona, October 2012

Stephanie Suhr

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

Building bridges

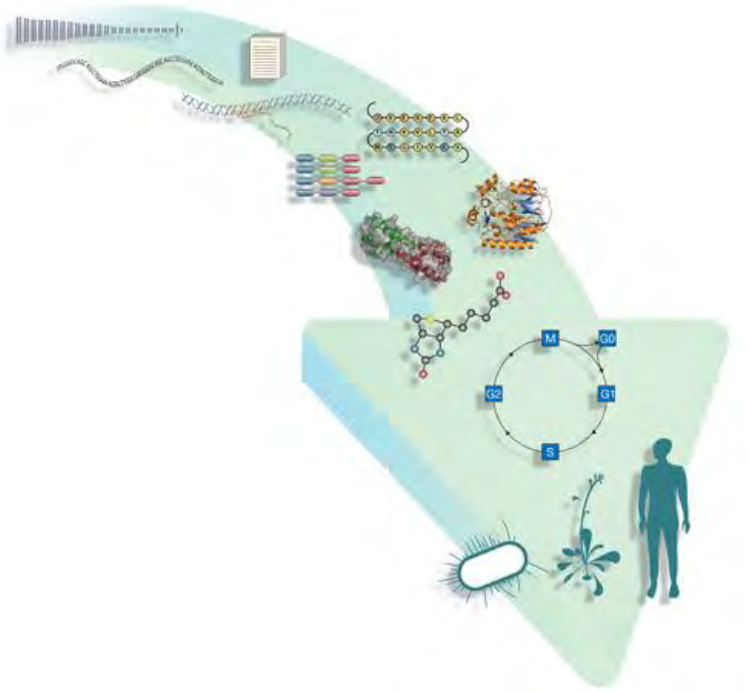
- FP7-funded cluster project
- 21 project partners in 9 countries
- BioMedBridges will bring together ten emerging Research Infrastructures in the Biological and Medical Sciences on the ESFRI roadmap
- RIs include biobanks, bioinformatics, translational research, marine resources, structural biology, mouse biology, imaging, clinical trials, highly contagious agents, and chemical biology
- Each of the ten RIs again has up to 50 partners that were or are involved in the preparatory phase!

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

The project in context: scope and possible reach

- The European Bioinformatics Institute coordinates BioMedBridges on behalf of



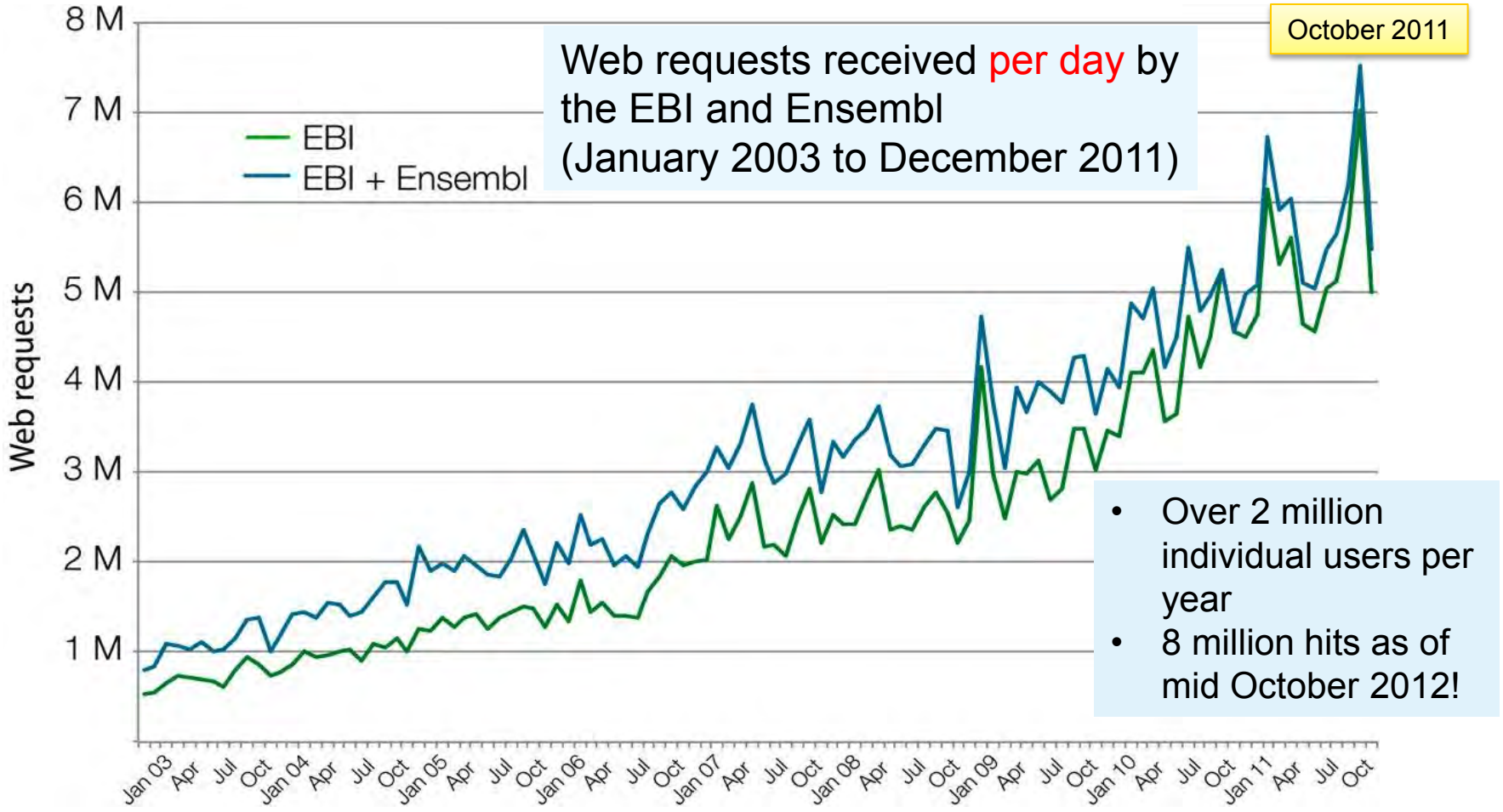
- EBI is Europe's hub for biological data services and research, supporting data deposition and exploitation
- Data resources include:
  - Genomes and functional genomics; nucleotide sequences
  - Chemogenomics; chemical entities
  - Protein sequences and activity
  - Macromolecular structures
  - Literature and ontologies
  - Pathways and systems

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

EBI users, 4 October 2012…

Web requests



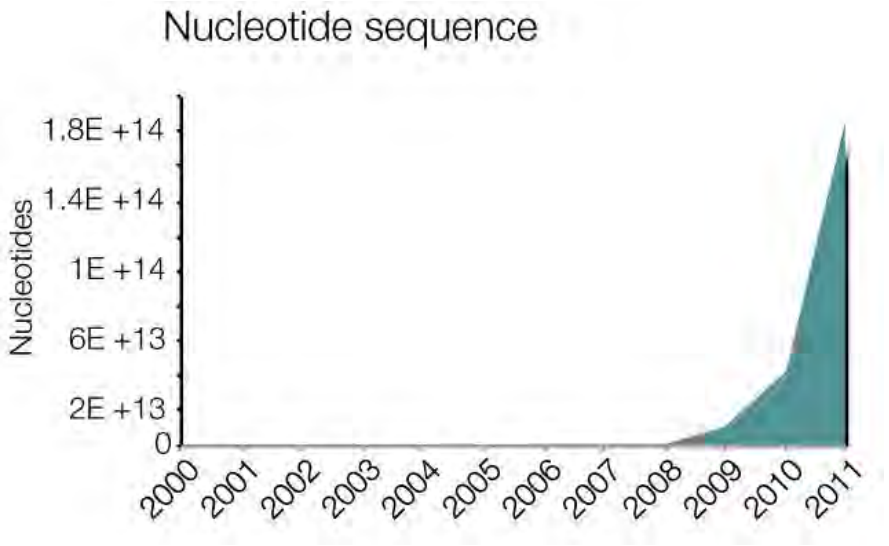
- Over 2 million individual users per year
- 8 million hits as of mid October 2012!

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

Biology is changing

- Data explosion
- New types of data
- High-throughput biology
- Emphasis on **systems**
- I/O is a bottleneck for many analyses



Growth of data at EMBL-EBI, example nucleotide sequences **(current total storage ca. 15 PB)**

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

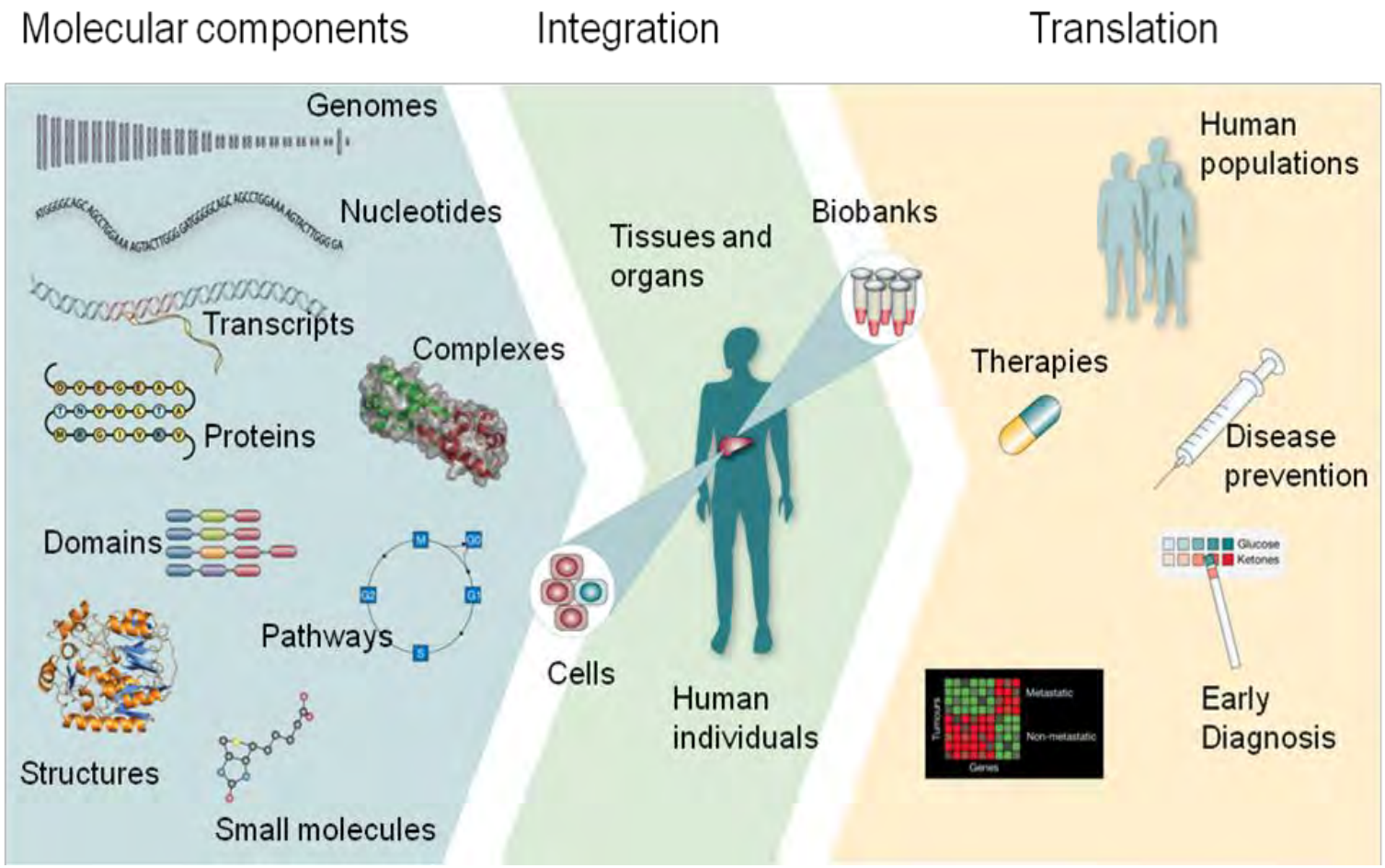
BioMedBridges objectives

- Computational „data and service" bridges between the BMS RIs
- interoperability between data and services in the biological, medical, translational and clinical domains
- Link basic biological research and data to clinical research and associated data.

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

From molecules to medicine…



BioMedBridges

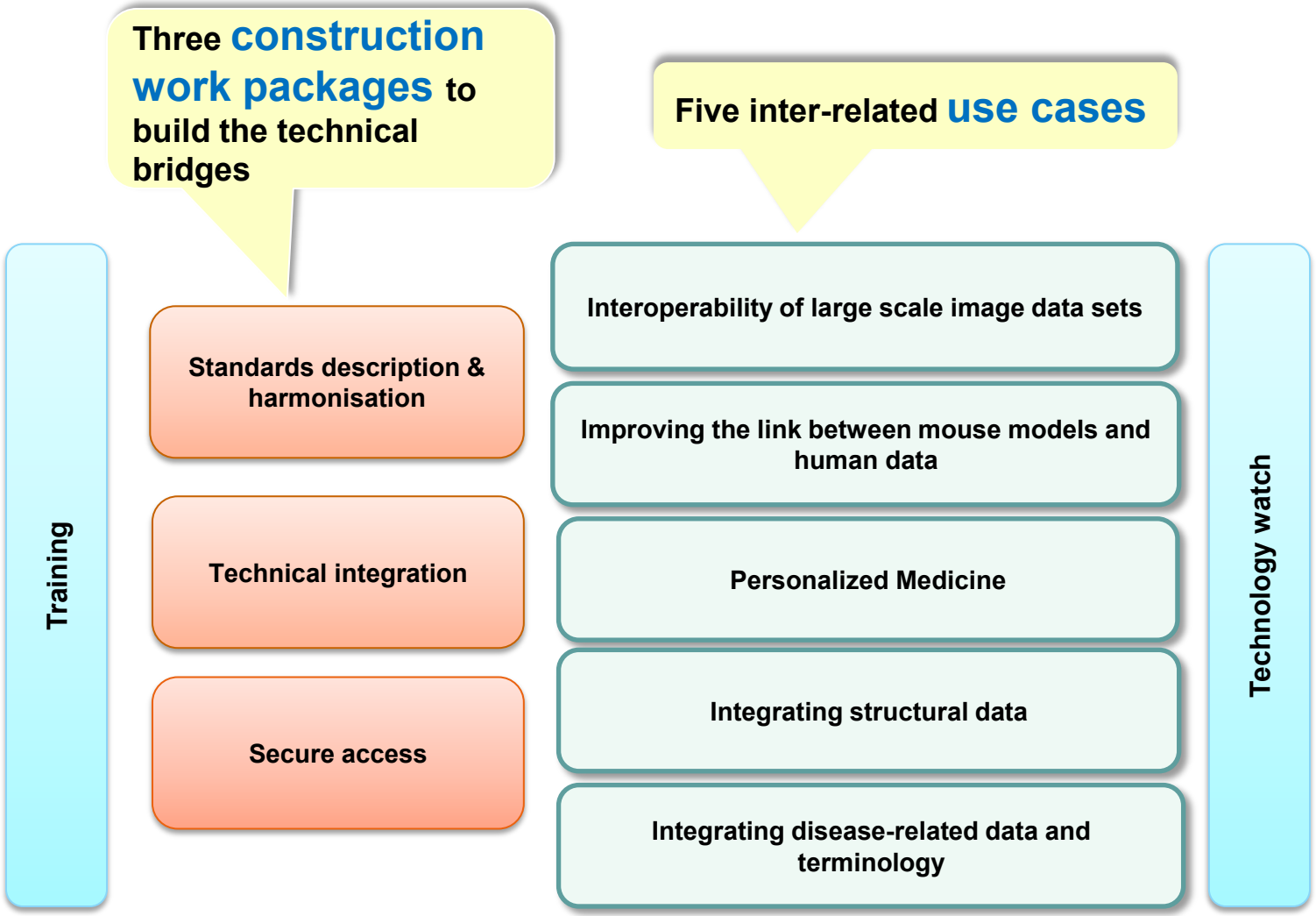
Multiple cross-overs between different communities!

SEVENTH FRAMEWORK PROGRAMME

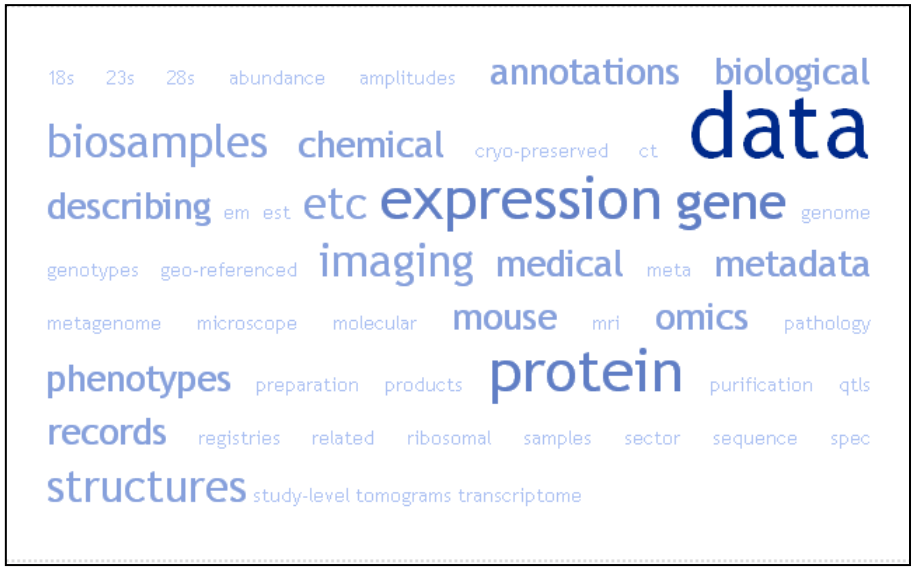
Beyond data availability and accessibility!

- "Building data and service bridges in the life sciences"
  - What is needed for scientific discovery?
  - What do researchers want/need?
  - What are the use cases and real world problems?
  - What data is involved?
  - What are the technical issues?

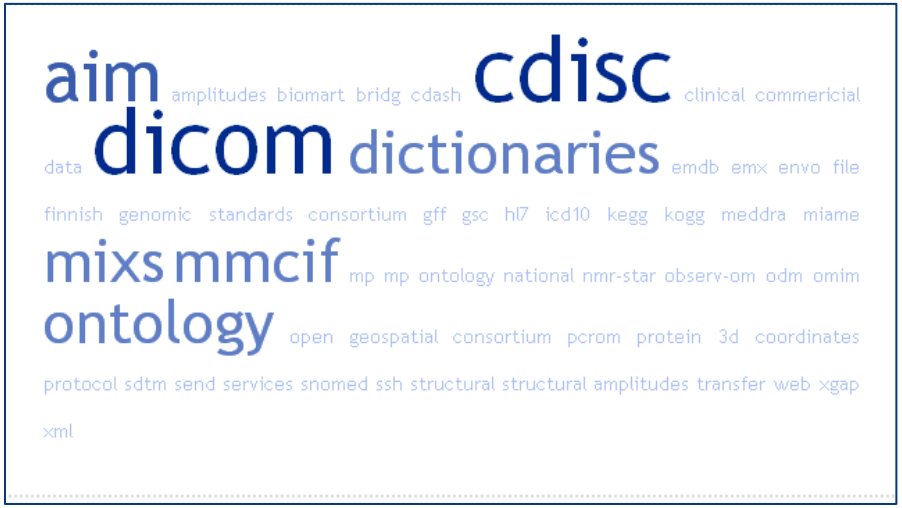
Active feedback between construction work packages and use cases to ensure services meet the needs of the scientific community



Heterogeneous data



Types of data



Data standards

Data formats

Research Infrastructures have different roles

- Survey of the ten participating RIs:
  - at least 80% of the respondents consume or produce/serve different types of data
  - at least 90% of them use standards when doing so
  - there are producer RIs and consumer RIs
  - consumer RIs also have preferred formats and standards to which they adhere
- Consumer RIs may (currently) be unable to share because:
  - there are restrictions concerning e.g. patient data or
  - the domain is not yet mature enough to have a portal or a central or federated repository

→ *More information: Deliverable 4.2 "Assessment of feasible data integration paths in BioMedBridges databases" (background, survey, technical approach)*

Current barriers to data integration

- Current barriers identified among the BioMedBridges partners fall into three different categories:

1. Sample/data standards and availability issues
2. Technical issues
3. People/ethical issues

Current barriers to integration

- Sample/data standards and availability issues
  - Lack of semantic interoperability: standards are poorly adopted in practice
  - Lack of a common disease ontology
  - Lack of common sample/data model (with joint annotation metadata, ontologies)
  - Lack of consistent data sets across various data domains relevant to translational research
  - Complementary and well-documented available public phenotyping data from mouse and man still need to be identified and brought together (example below)
  - Lack of well-organised public domain data on drugs-target pairs
  - Lack of structural data for many proteins, i.e. much required data simply doesn't exist
  - Lack of comprehensive genome variant disease association information in the public domain in an easily usable manner
- Technical issues
  - Lack of suitable portals / data displays / queries / standardized method to define data mappings for pooling of data from different resources
  - Lack of adoption of standards among software packages
  - Lack of user friendly tools for organising, managing and sharing data
  - Lack of privacy/security
- People/ethical issues
  - Language barrier
  - Crossing a skill boundary - few people can interpret both sorts of data
  - Reluctance to share biological/clinical data

Standards description and harmonization

- To link data between different domains, they must use:
  - common identifiers
  - harmonised content, syntax and semantics
- BioMedBridges will create:
  - an online dictionary of common molecular identifiers of the BMS research infrastructures
  - a registry of standards used
  - a meta service registry

Technical integration

- Data will be made accessible/linkable:
  - via use of REST-based Web-Services interfaces optimised for browsing and for programmatic access
  - by exposing appropriate meta-data information via use of Semantic Web Technologies

- BioMedBridges will pilot the use of semantic web technologies in high-data scale biological environments

Ethics requirements and data protection

- Access to much of the data in the biomedical sciences has Ethical, Legal or Societal Implications (ELSI)
- BioMedBridges is:
  - working with these across organisational boundaries
  - working with these across national boundaries
  - linking these to other (types of) data
- This has implications for:
  - data protection and data security
  - informed consent by study participants/data subjects
- Data made available within BioMedBridges will *not* include Personally Identifiable Information.

Secure access

- BioMedBridges will build a security framework that:
  - Will address the ethical, legal and regulatory issues resulting from sharing data and providing access to biomaterials
  - is in compliance with national and European regulations, privacy rules and access requirements

Use case: PhenoBridge – crossing the species bridge between mouse and human

- Fundamental differences in the terminologies used by each community to describe the same phenotypes in mouse and human
- Develop common standards and ontologies for comparable diabetes and obesity-related large-scale datasets in mouse and human to bridge the phenotype gap

→mouse models more useful: direct translation of different conditions between the two species

→opens the use of extensive, existing mouse phenotype data resources by clinical researchers.

BioMedBridges

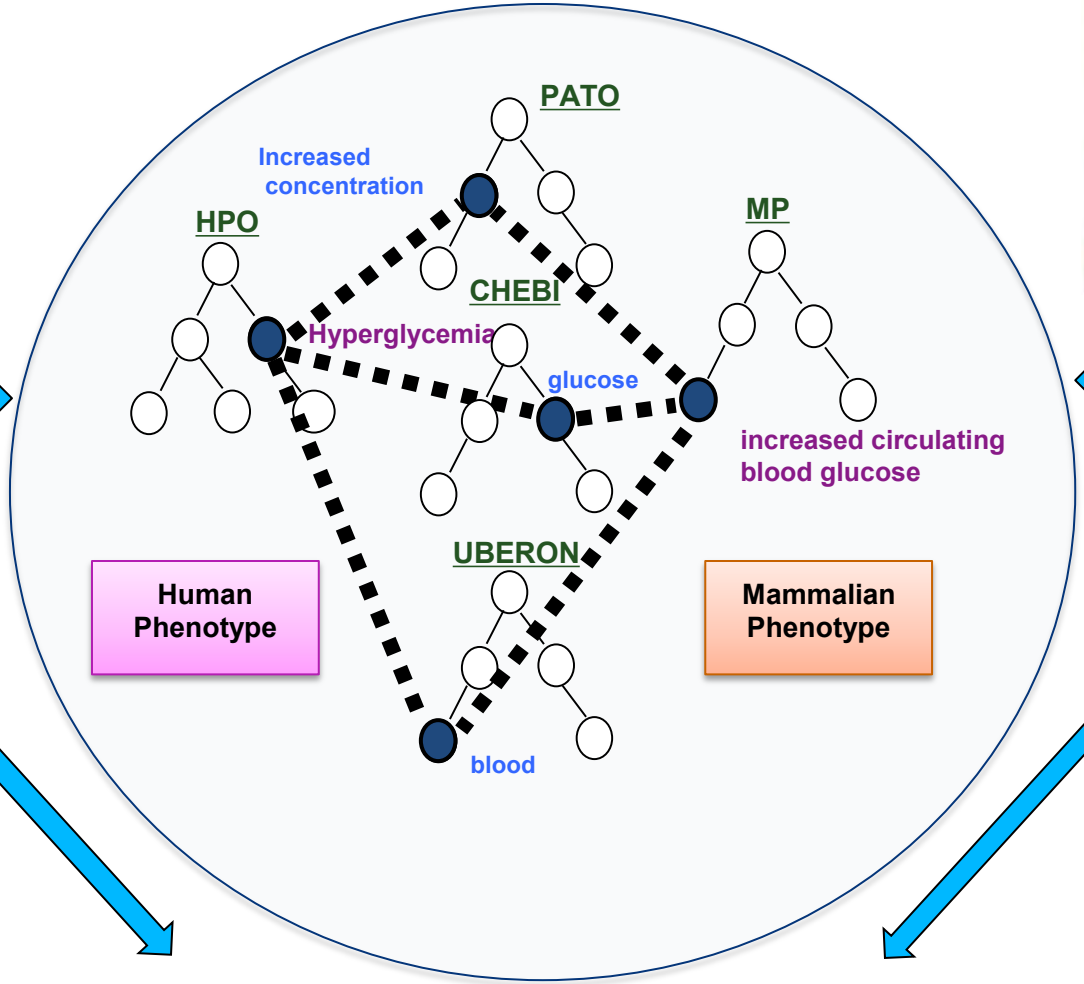
elixir

BBMRI
Biobanking and Biomolecular Resources Research Infrastructure

SEVENTH FRAMEWORK PROGRAMME

Human Data

Over 100 GWAS studies annotated to "Diabetes"



Model Organism Data

Over 500 Mouse genotypes annotated with "increased circulating blood glucose"

Adapt a tool* that will find matching phenotypes in the other species

- → **Novel disease candidate**
- → **Novel pathways**
- → **Novel detection assay**
- → **Novel therapies**

*MouseFinder http://mousemodels.org/

BioMedBridges is about:

- Adding value to existing data by linking it
  - Creating links between available data that were not linked before will hugely increase the potential for new discoveries
- Bringing together different communities in the biological and medical sciences
  - create a common understanding of and approach to data (standards, formats, etc.... and how to make it linkable!)
- Turbo-charging the participating ten new biomedical sciences research infrastructures as they build up new resources

BioMedBridges

SEVENTH FRAMEWORK PROGRAMME

Thank you for your attention.