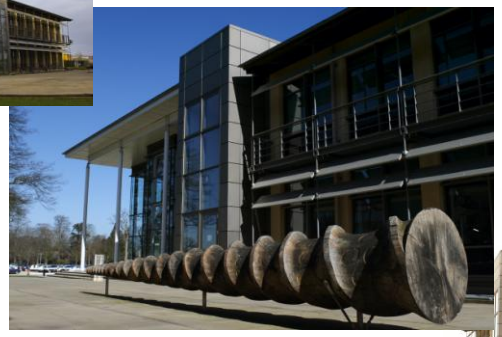
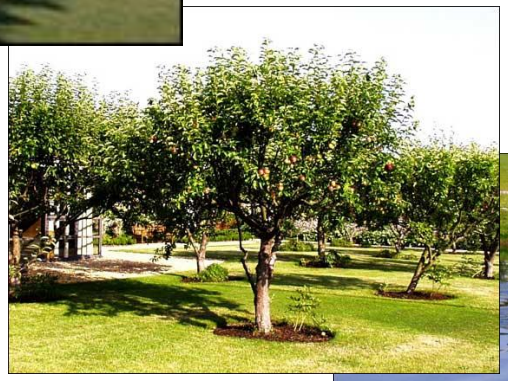


Curation in biology: Past, Present and Future



Janet Thornton
EMBL-EBI
April 2009

Overview

- Introduction: Personal experience with the PDB
- Curation @ EBI
- Future of Biological Information in Europe – ELIXIR
- Challenges for Curation

Overview

- Introduction: Personal experience with the PDB
- Curation @ EBI
- Future of Biological Information in Europe – ELIXIR
- Challenges for Curation

The Protein Data Bank (PDB)

- PDB – the ‘oldest’ biological data resource, established in early 1970s
- Distributed on large magnetic tapes, sent by post
- Format of entries ‘fixed’, and based on Fortran requirements (80 characters/line)
- Keywords on every line
- Depositors were ‘computer-literate’

Challenges in PDB in Early 90s

- Manual data entry
  - Nomenclature &Typos (26 ways to spell E. coli)
- Lack of standard ontology
  - No consistent way to handle modified amino acids
- Ligand chemistry poorly defined – lack of standards
- Inconsistencies – no verification
- Wrong structures – lack of validation
- Inefficient - it took several days to deposit

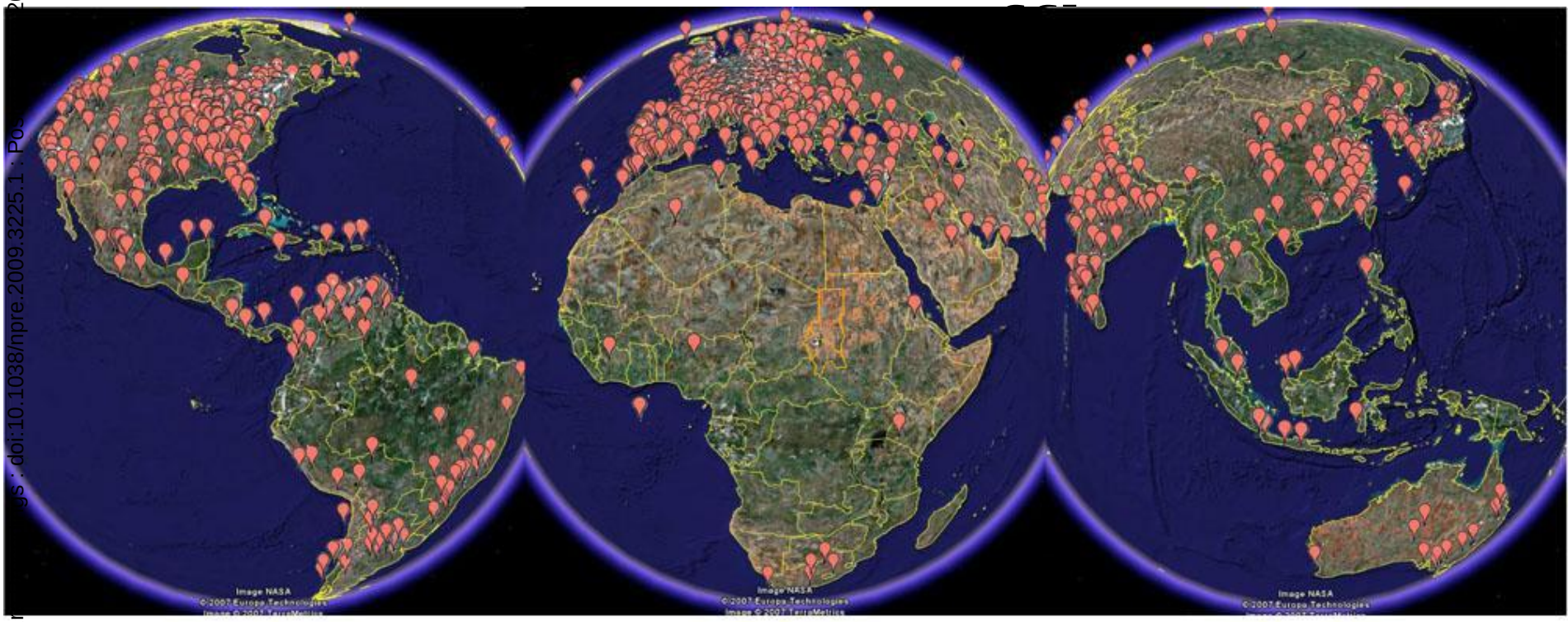
The Reluctant Curator!

- Initially by hand! (grouping similar proteins together)
- Then more automatically
  - Comparisons in 3D – SuperA & SuperB
  - Validation Tools (PROCHECK)
  - Classification (CATH) + Manual Curation
  - Motif/Template Definition in 3D
  - 3D Searching
- BUT – functional annotation still a challenge!

Worldwide Protein Data Bank

www.wwpdb.org

Natur... gs : doi:10.1038/npre.2009.3225.1 : Po... 2009



wwPDB today

- Manual data entry – **now limited**
  - Nomenclature & Typos (26 ways to spell *E. coli*) - **now external reference files**
- Lack of standard ontology
  - No consistent way to handle modified amino acids - **now Ontologies used**
- Ligand chemistry poorly defined - **now remediated**
- Inconsistencies – no verification - **now sophisticated verification suite applied to each entry**
- Wrong structures – lack of validation - **now validation auite to check reliability of data**
- Inefficient - it took several days to deposit a structure **now web interface – but currently being redesigned!**
- Natural complexities of data led to inconsistent entries **now rules for most eventualities but still some complexities**
- Lack of functional Information (metadata/sample data) **Still a problem – see later**
- No tools to search PDB at any level **now many tools**

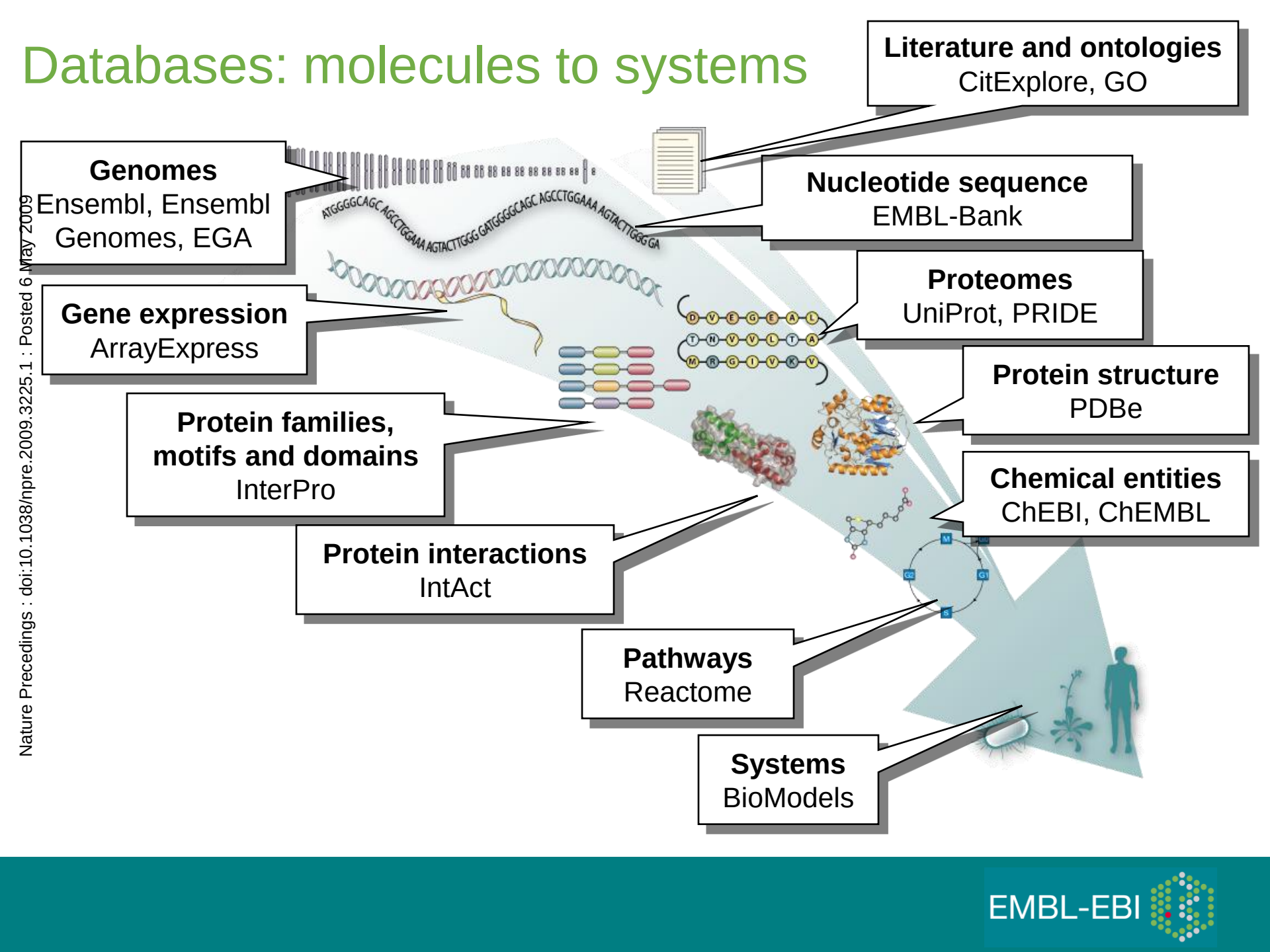
My conclusions

- For data resources to be useful, careful curation is vital
- Curation and Software development go hand-in-hand
- Automation where possible is essential
- Investment in good submission software pays off
- Manual curation remains essential, but the goal should be to restrict this to those parts of the process, which cannot yet be handled automatically
- If the data are well curated, this will save many scientists time and money (& frustration!)

Overview

- Introduction: Personal experience with the PDB
- Curation @ EBI
- Future of Biological Information in Europe – ELIXIR
- Challenges for Curation
  - Flood of Sequence data
  - Curating Biological function
  - Heterogeneous Data

Databases: molecules to systems



CURATION @ EBI

Literature and ontologies
CitExplore, GO

Genomes
Ensembl, Ensembl Genomes
HGNC **5**

Nucleotide sequence
ENA **6**

Gene expression
ArrayExpress **~6**

Proteomes
UniProt **13**, PRIDE **2**

Protein structure
PDBe **6**

Protein families, motifs and domains
InterPro

Chemicals
ChEBI **2.5** & ChEMBL **2.0**

Protein interactions
IntAct **~2**

Pathways
Reactome **3**

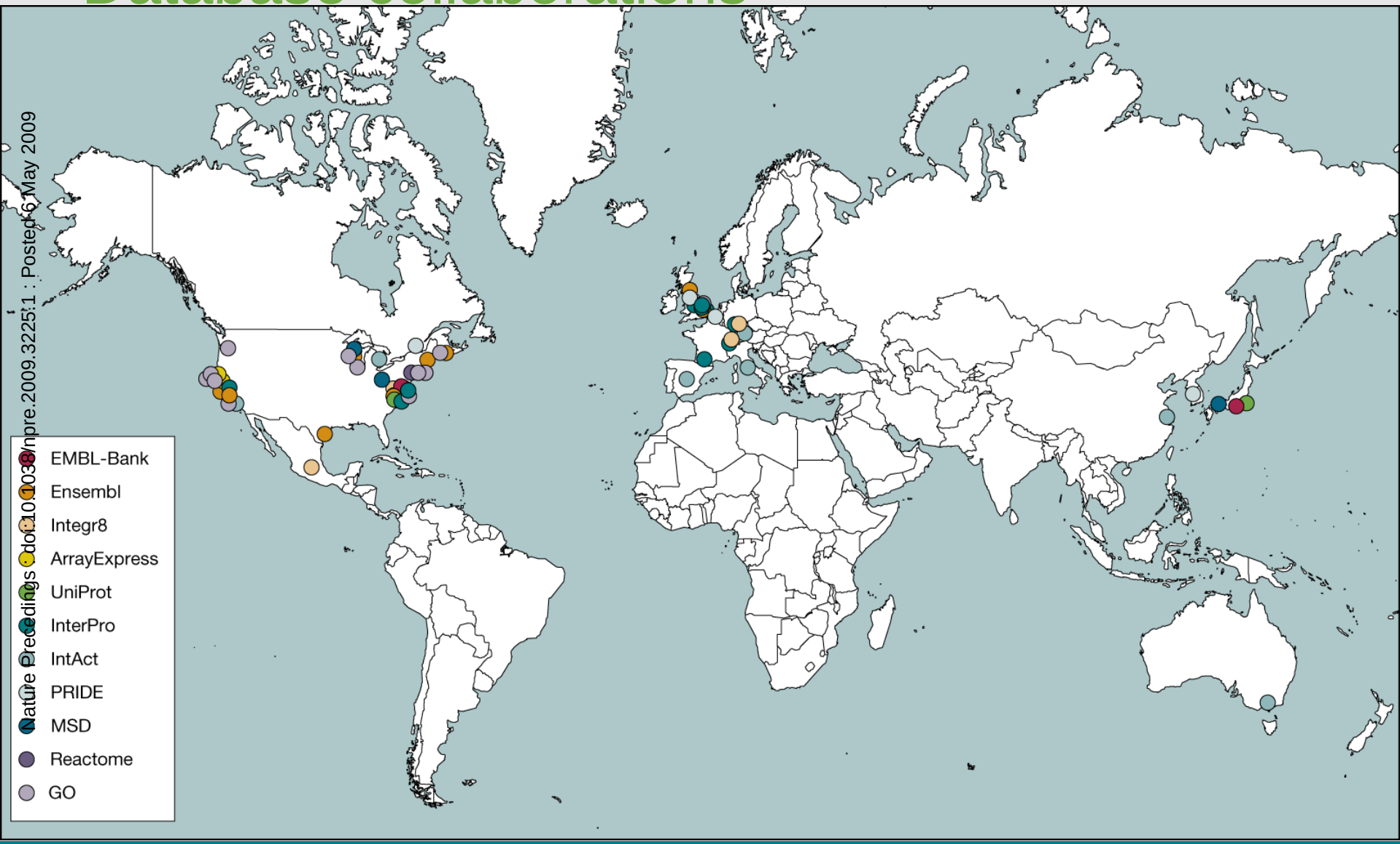
Systems
BioModels **1**

>50 Curators in total

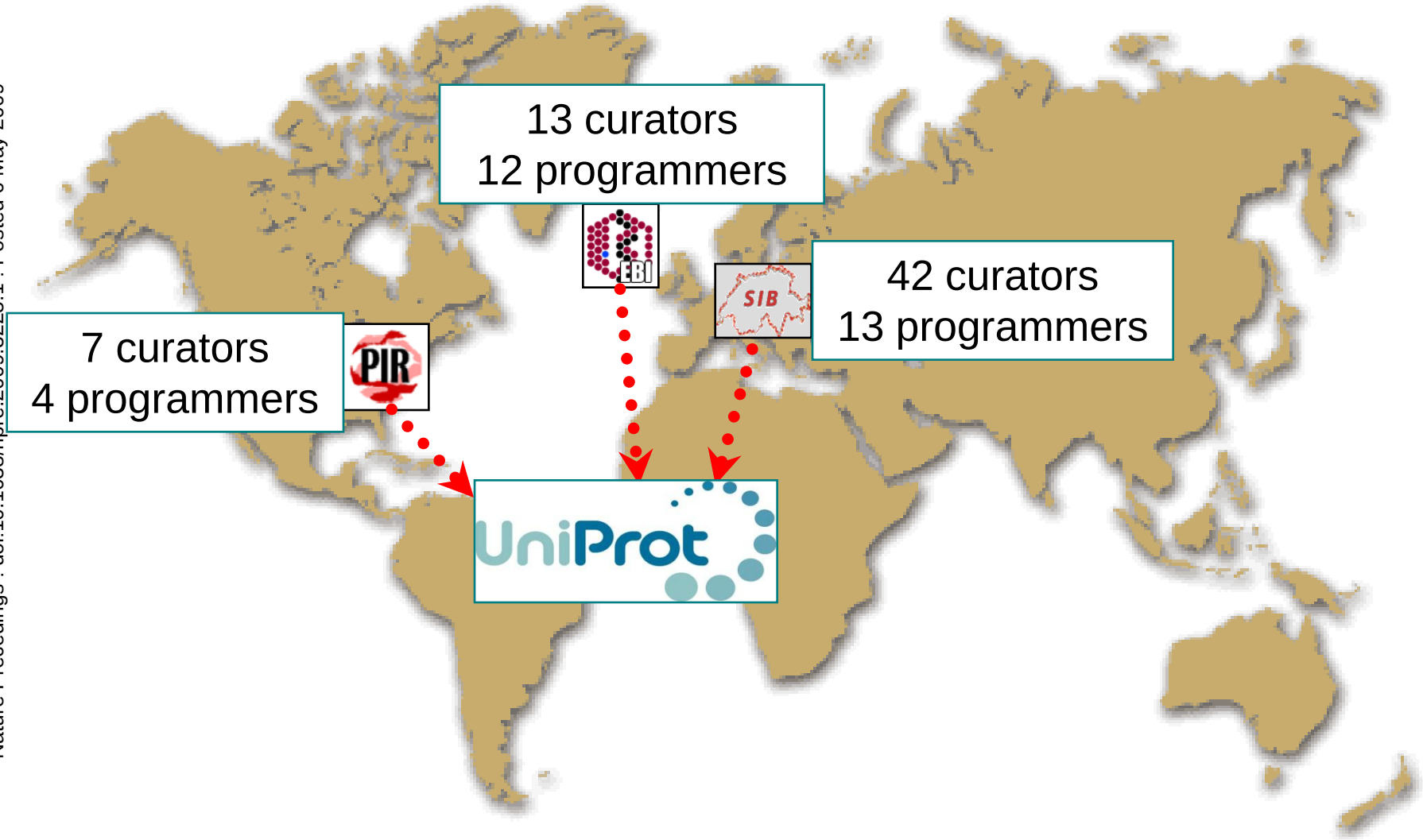
Nature Precedings : doi:10.1038/npre.2009.3225.1 : Posted 6 May 2009

EMBL-EBI

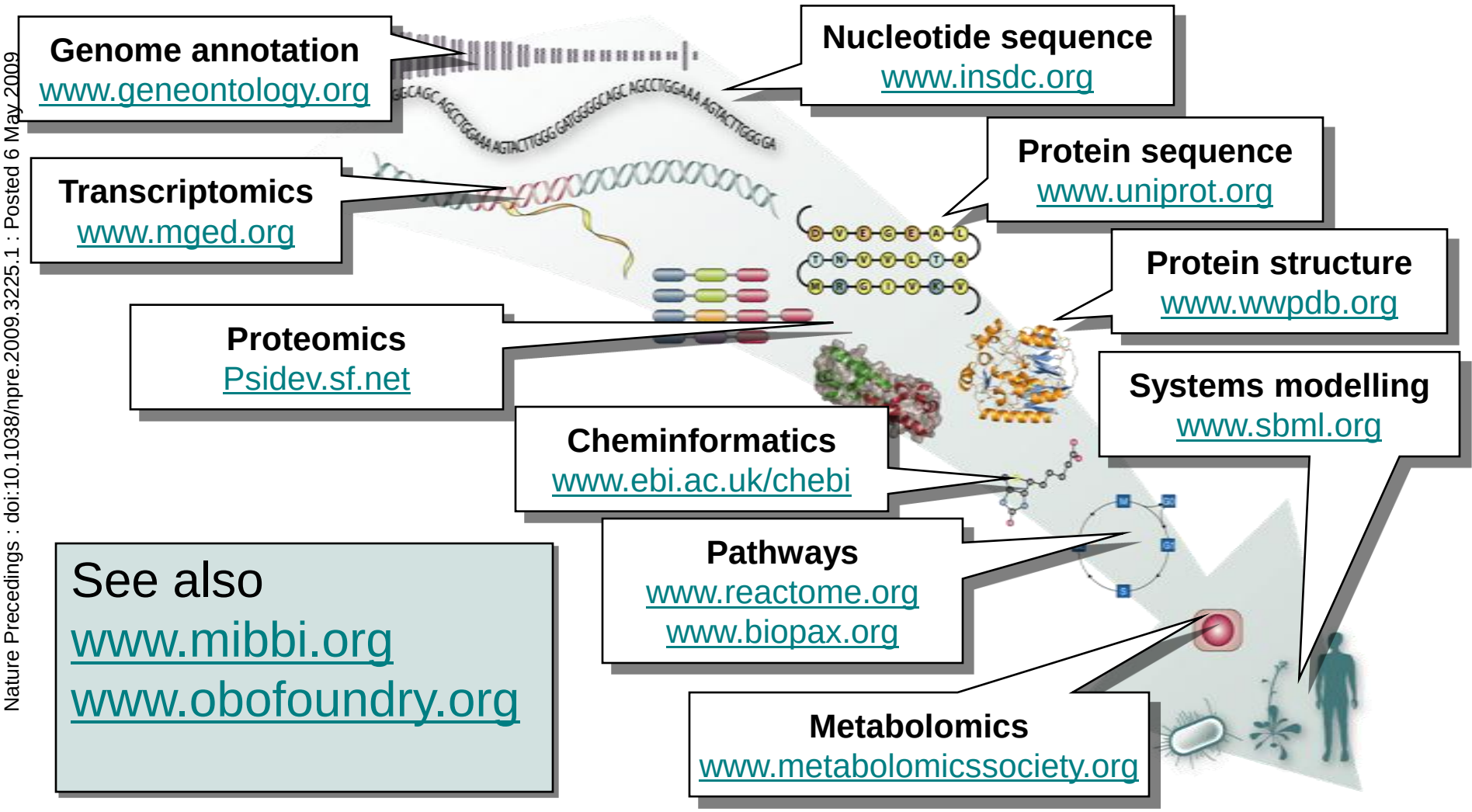
Database collaborations



1. UniProt Consortium



Standards development



Nature Precedings : doi:10.1038/npre.2009.3225.1 : Posted 6 May 2009

Curation tasks

- Curation and updates of new data
  - Verification/Quality Control/Validation
- Maintenance of cross-referencing to related databases
- Assignment of unique approved symbols and names
- Submitter and user support
  - Maintenance of documentation for users
  - Training

- Collaboration with ‘partner’ databases eg in wwPDB

- Development and implementation of standards for new data representation and annotation
  - Collaboration with groups working in associated areas to develop standards, ontologies and mappings

Different Types of Curation

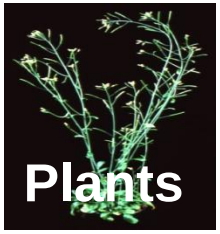
- Data Submission
  - ENA; UniProt; ArrayExpress; wwPDB; ChEBI; PRIDE
- Value Added Curation
  - UniProt Knowledge Base
  - GOA
  - Many specialist data resources

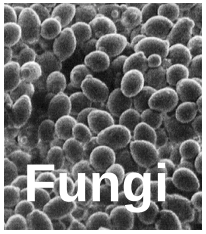
PDBe Curation Team Tasks

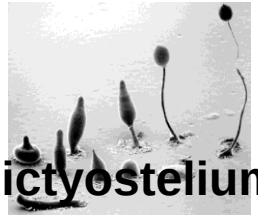
- Annotation of depositions into the Protein Databank (PDB) and the Electron Microscopy Databank (EMDB).
- Maintenance and development of Deposition and Annotation Software and workflow pipelines.
- Improvement in data format and overall data integrity and consistency in collaboration with other wwPDB partners (RCSB, PDBj and BMRB).
- Outreach and training activities in the form of roadshows and courses at EBI and externally.

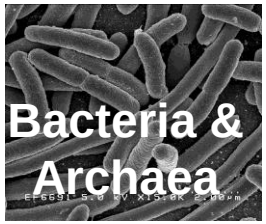
Uniprot Curation tasks

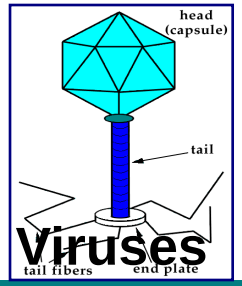
- **Species-specific curation**











- **Other priority areas**
  - Post-translational modifications
  - 3D-structural data
  - Enzymes
  - Protein interactions
  - Integration of data from large-scale proteomics studies
  - Peptide submissions
  - Journal scanning for unsubmitted peptide sequences
  - Development of controlled vocabularies
  - GO annotation

PRIDE (Proteomics Data) Current Curation tasks

Limited to data submission support

- Mass spectrometry derived data: very heterogeneous nature in terms of experimental approaches, instrumentation, data formats,…

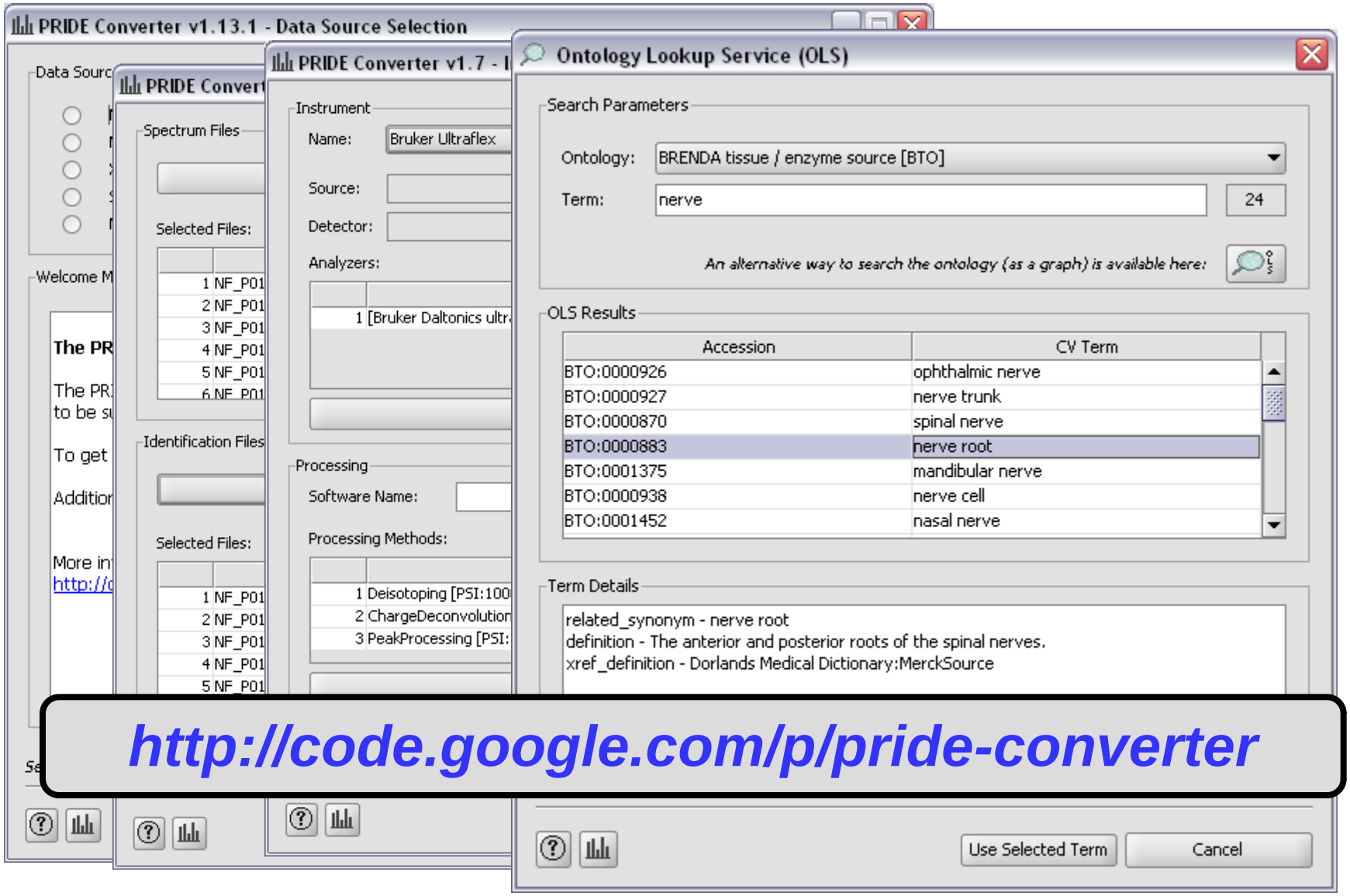
- The format needed is PRIDE XML.

- Conversion of proteomics data to PRIDE XML format can be time difficult and very time-consuming (especially for pure biologists).

Some tools are now available to ease the process

- Proteome Harvester: suitable only for small scale submissions.
- PRIDE Converter (http://code.google.com/p/pride-converter)

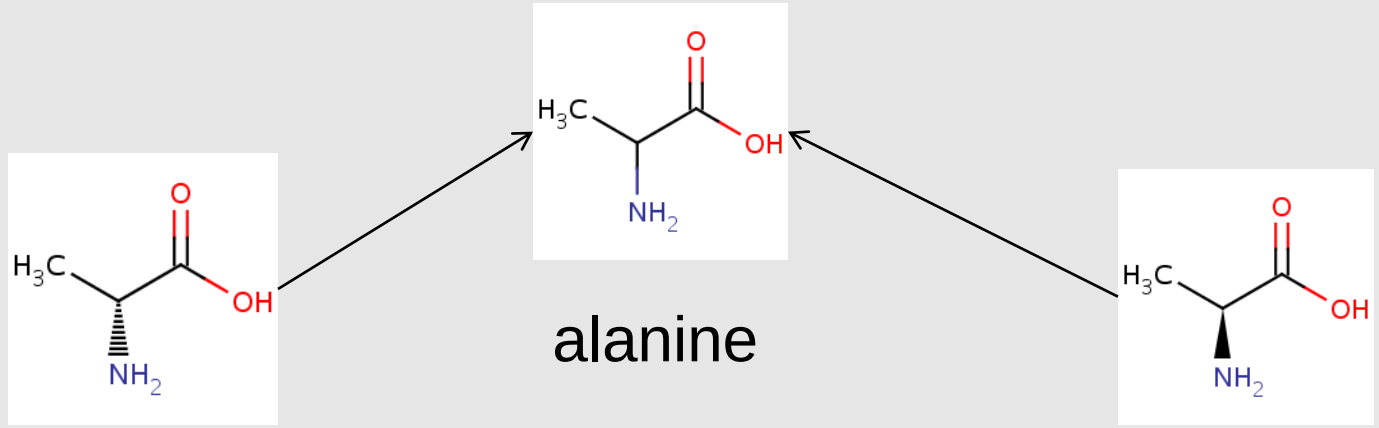
The PRIDE Converter tool to create submissions



http://code.google.com/p/pride-converter

ChEBI Curation Tasks

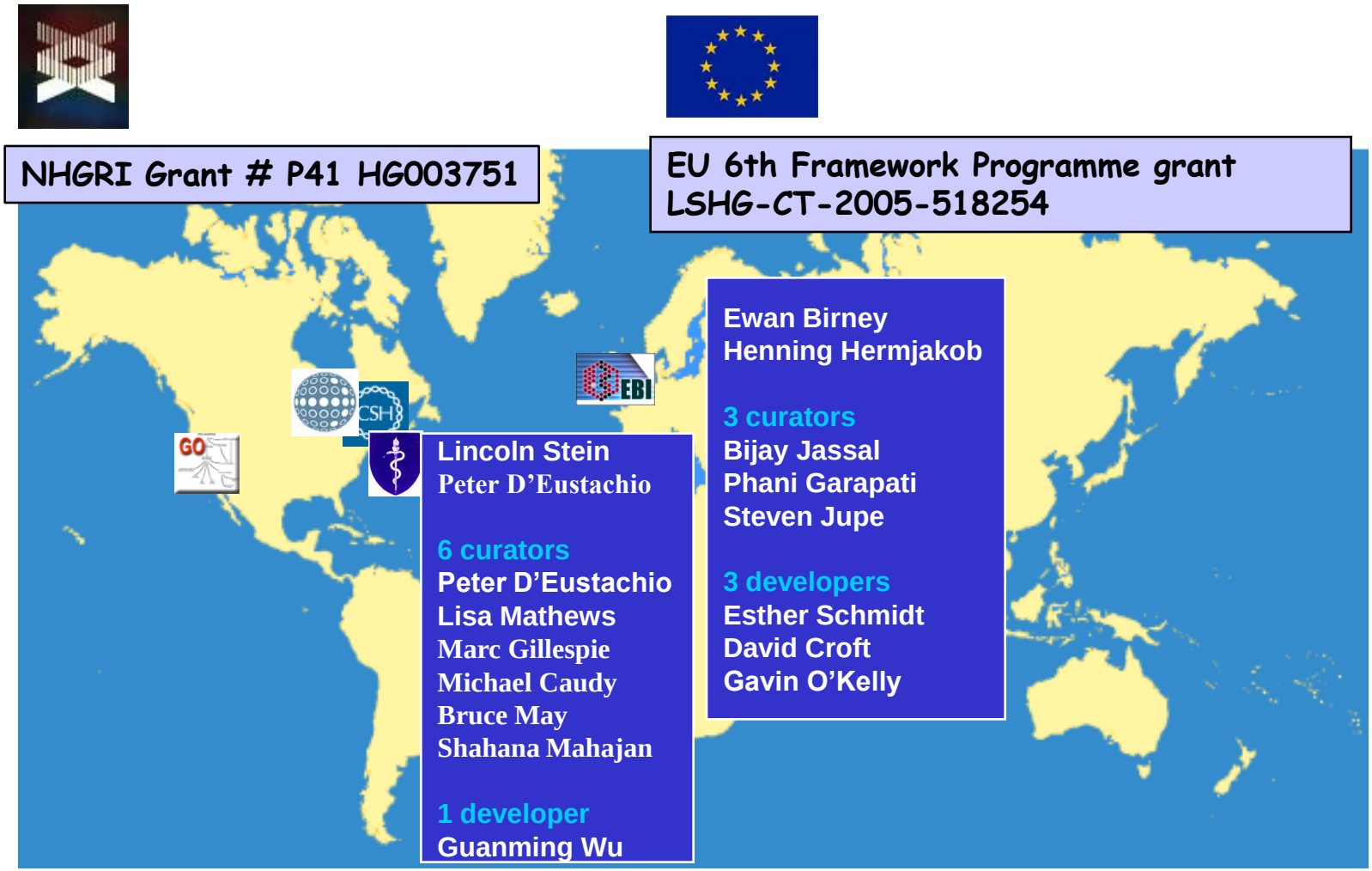
- Nomenclature disambiguation
- Classification of chemical entities within an ontology
- Annotation of chemical structures
- Dictionary annotation (new terminology)



Reactome

The **Reactome** project is a collaboration among Cold Spring Harbor Laboratory/New York University, The European Bioinformatics Institute, and The Gene Ontology Consortium to develop a curated resource of core pathways and reactions in human biology. Reactome is a free on-line resource, and Reactome software is open-source. We have **9 curators** and **4 developers** working on the project.

Nature Precedings : doi:10.1038/npre.2009.3225.1 : Posted 6 May 2009



Curation Tasks

- Human-centric events are authored by biological researchers or curators with expertise in their fields
- Information is maintained by the Reactome editorial staff and is peer-reviewed before public release
- Cross-referenced with the sequence databases at NCBI, Ensembl and UniProt, the UCSC Genome Browser , HapMap, KEGG (Gene and Compound), ChEBI, PubMed and GO.
- High quality curation with a small team – current coverage ~17% of SwissProt

Current and Future Challenges I

- DATA
  - Ongoing growth data
  - New sorts of data eg new sequencing data
  - Maintenance of historic data
  - Merging manual and automatic data
  - Improving verification and validation (QC)

- STANDARDS
  - Connecting with external groups to drive standards development

- LINKS
  - Creating rational connections both within database and outward to external data

Current and Future Challenges II

- INTERFACE TO USERS
  - Improving presentation of data to users
  - Improving submission systems to benefit submitters, and increase curation efficiency and throughput
  - Ontology driven GUIs
  - ImprovingTraining

- ADDING VALUE
  - Improving automatic annotations & their coverage
  - Adding semantic value
  - Evidence attribution

- ETHICAL CONSIDERATIONS

Data Submission

- Keys to success & efficient deposition
  - Well defined ontology & standards
  - Sufficient, but not too much, detail
  - Basic reference files for standard information
  - Good format (eg XML)
  - Robust & powerful tool for submission
  - Data harvesting
  - Easy to use interface
  - Get depositor to do as much work as possible – but make it easy for them

Value Added Curation

- Keys to Success
  - Domain Knowledge
  - Careful Curation
  - Well defined ontology & standards
  - Sufficient, but not too much, detail
  - Robust & powerful tools to aid curation
    - Data harvesting
    - Easy to use interface for curated data input
  - Powerful tools to search the literature

Overview

- Introduction: Personal experience with the PDB
- Curation @ EBI
- Future of Biological Information in Europe – ELIXIR
- Challenges for Curation
  - Flood of Sequence data
  - Curating Biological function
  - Community Annotation

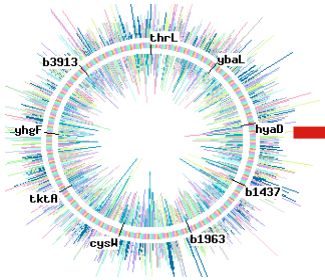
Consolidating Infrastructure for Biological Information in Europe

EMBL-EBI

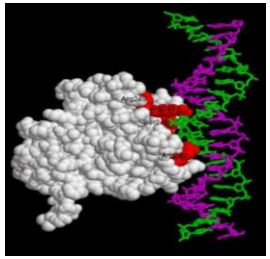
EBI is an Outstation of the European Molecular Biology Laboratory.

ELIXIR EUROPEAN LIFE SCIENCES INFRASTRUCTURE FOR BIOLOGICAL INFORMATION

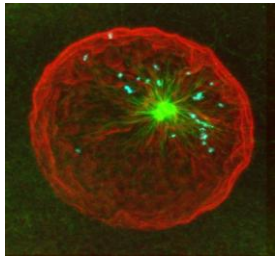
A European Infrastructure for Biological Information



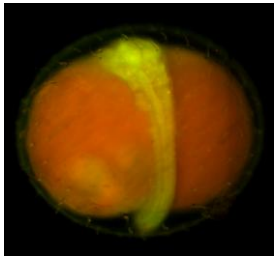
Genome



Protein



Cell



Embryo

Fruitfly

Mouse

Human
Development,
Ageing, Disease

What is Elixir?

- An EU Framework 7 Preparatory Phase Project
- 32 member consortium engaging many of Europe's funding agencies and research institutes
- Deliverable is memorandum of understanding between partners for the implementation phase
- Elixir Website: www.elixir-europe.org

ELIXIR Mission

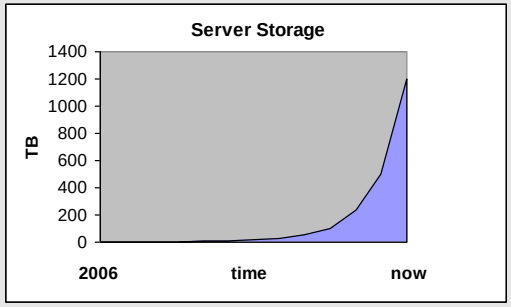
To construct and operate a sustainable infrastructure for biological information in Europe, to support life science research and its translation to medicine and the environment, the bio-industries and society.

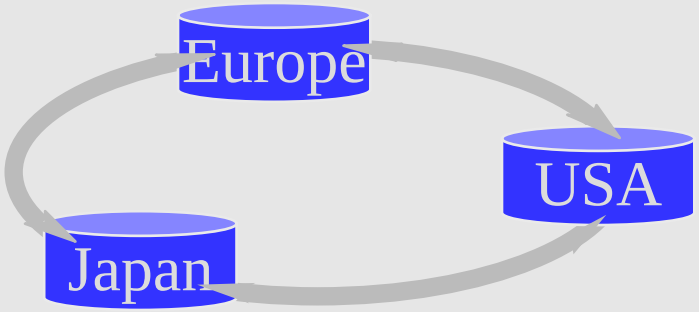
This will contribute to improvements in all human endeavour associated with living systems including:

- **health and medicine**
- **the environment**
- **Agriculture**
- **Fisheries**
- **Forestry**
- **Biotechnology**

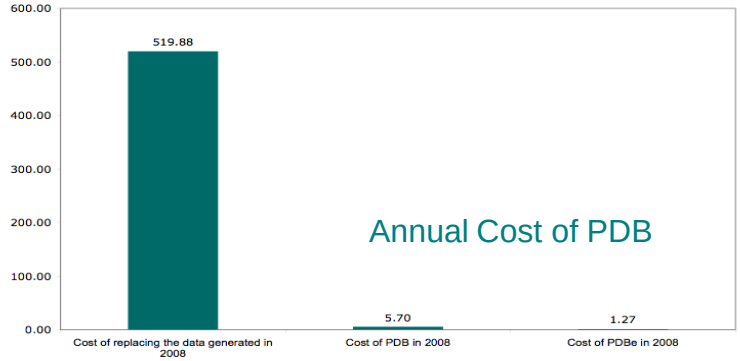
Why do we need ELIXIR?

- Data Growth
- Global context
- Very large user community:
  - 3.3 m web hits/day
  - 20,000 unique users per day
- Value for Money
- Need for increased funding
- Impact on medicine & agriculture
- Impact on society & bioindustries





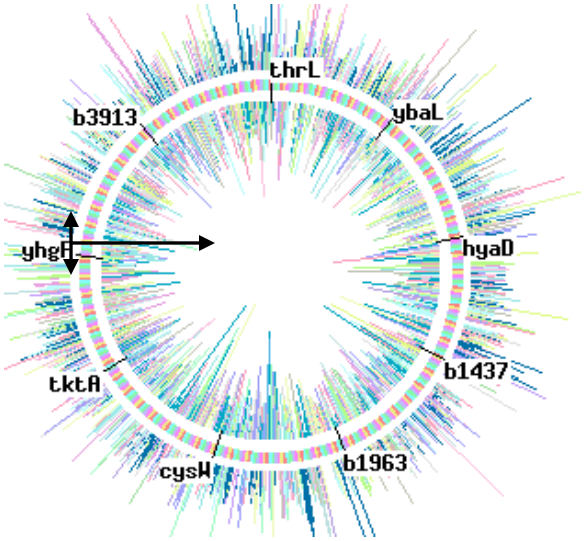
Data collection in 2008



<3%

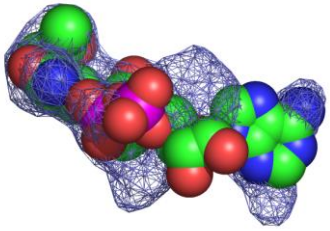
EMBL-EBI

Integration with Other Data is increasingly important
e.g. Linking from Molecules to Medicine & Agriculture



Genomes

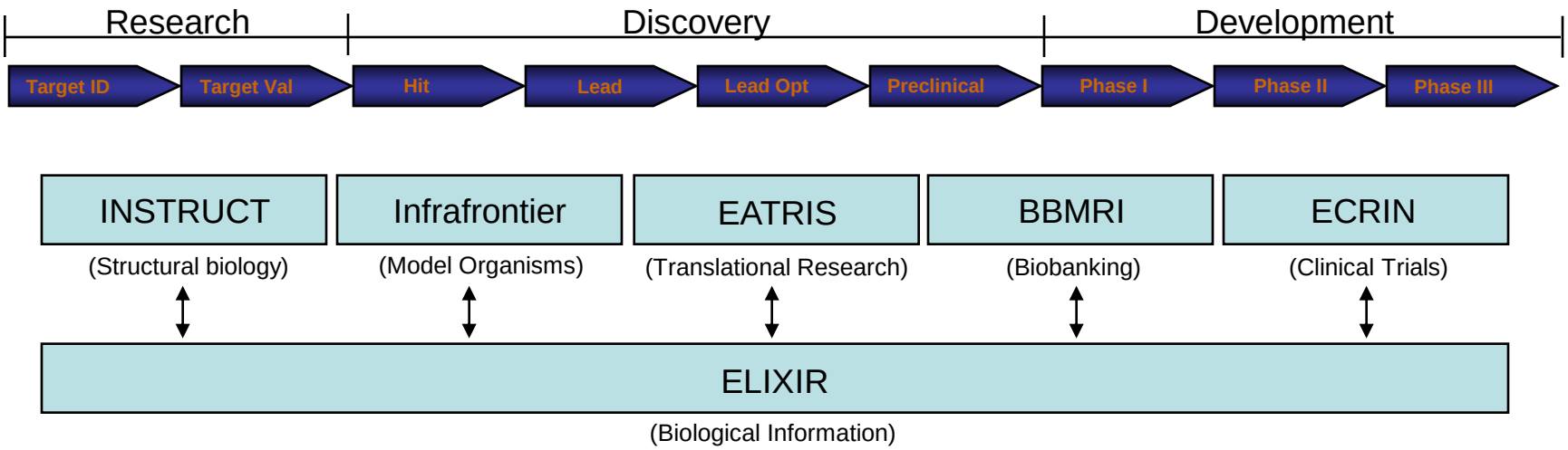
Proteins



Metabolites

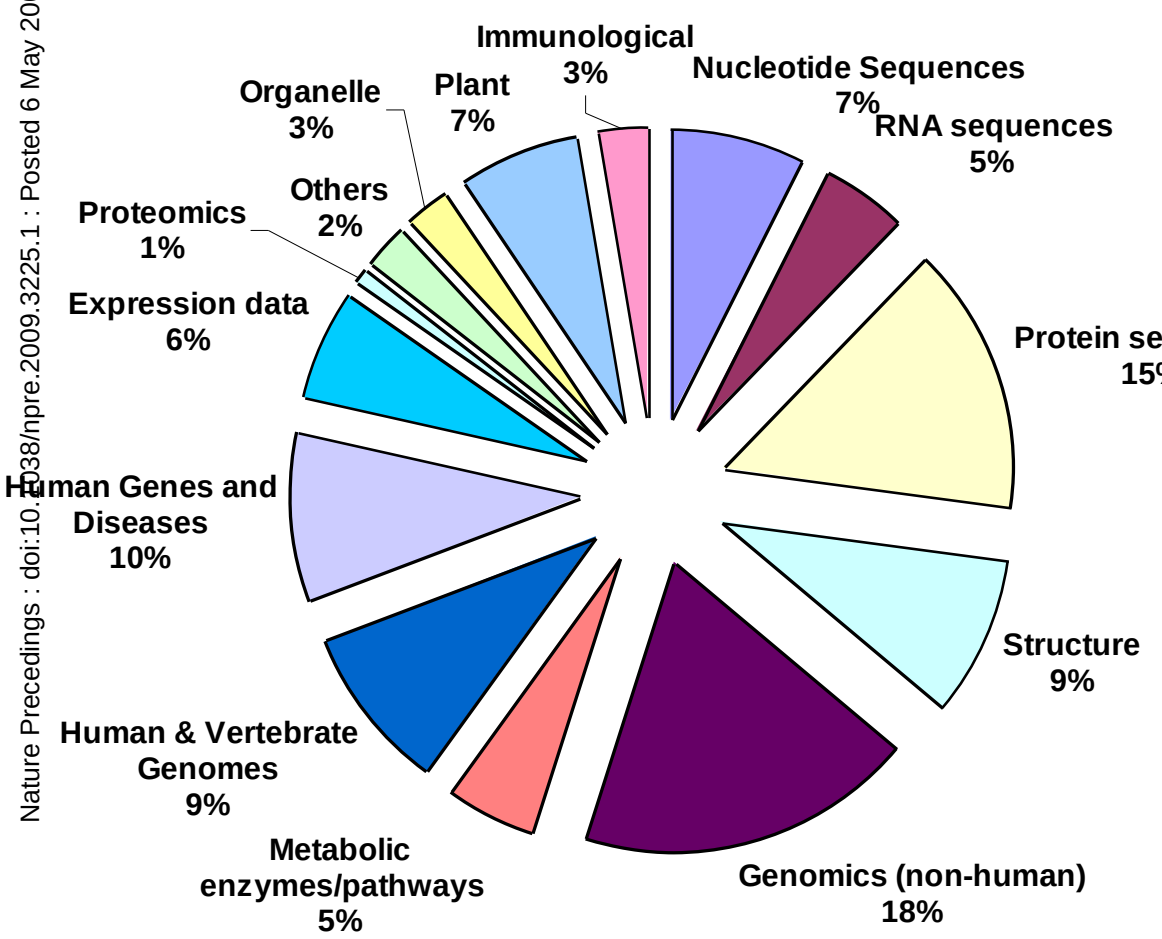
ESFRI Biology RI proposals.

INSTRUCT	Integrated Structural Biology Infrastructure	www.strubi.ox.ac.uk
Infrafrontier	Infrastructure for Phenomefrontier and Archivefrontier	www.emma.rm.cnr.it
EATRIS	The European Advanced Translational Research Infrastructure	http://www.eatris.eu/
BBMRI	European Biobanking And Biomolecular Resources	www.biobanks.eu
ECRIN	Infrastructures For Clinical Trials And Biotherapy	www.ecrin.org
ELIXIR	Upgrade Of European Biological Information Infrastructure	www.elixir-europe.org



ELIXIR: *a **sustainable** infrastructure for biological information in Europe*.

Specialised Molecular Data Resources



More than 700 in total

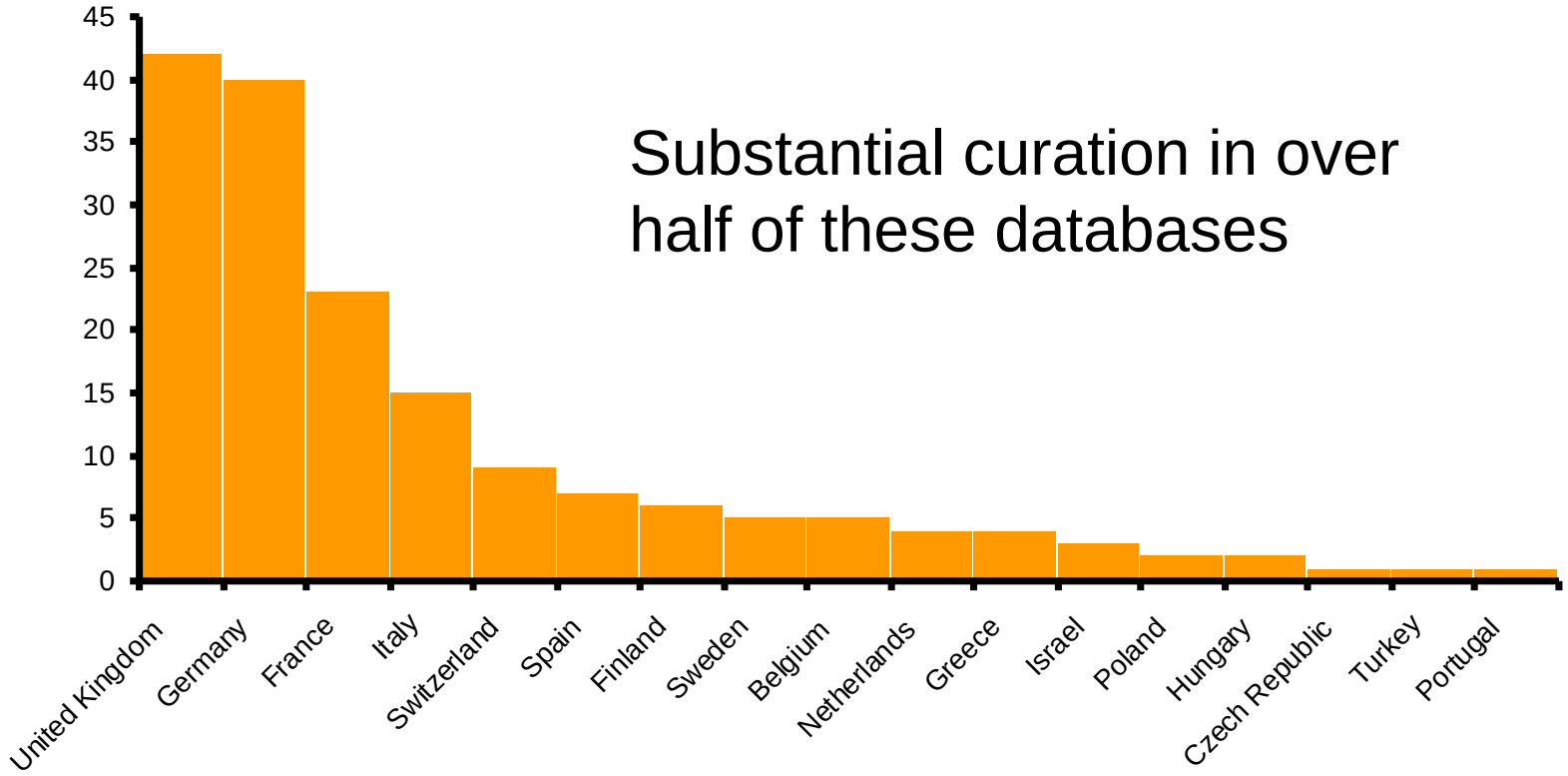
30% in Europe

(All use core resources as reference data)

Galperin (2005 NAR)

Number of Databases in Europe:

Response to ELIXIR Survey - 170 out of 508 URLs



Substantial curation in over half of these databases

Rationale for ELIXIR

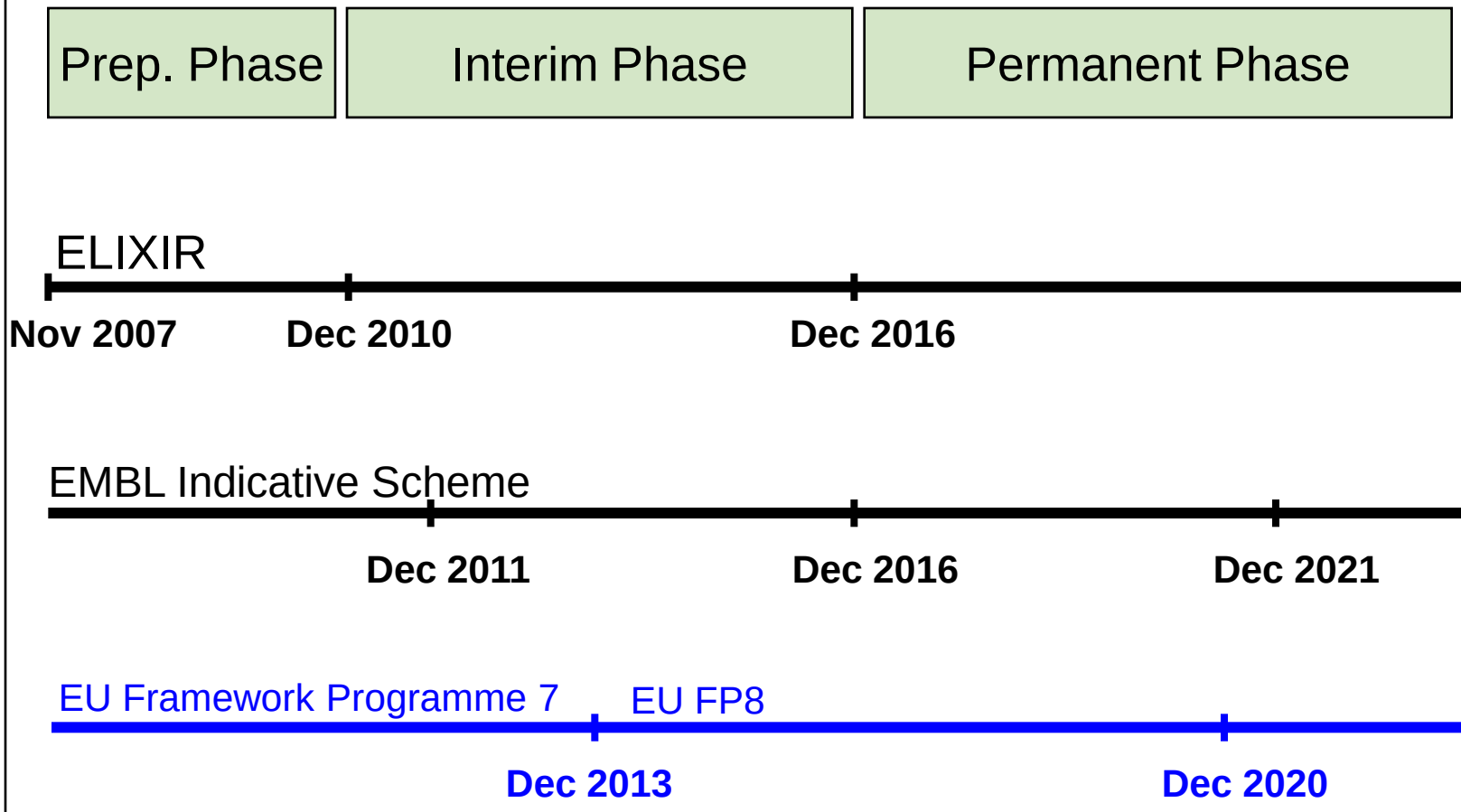
- **Optimal Data Management**
  - **Coordinated Data Resources with improved access**
  - **Integration and interoperability of diverse heterogeneous data**
- **Forge Links to data in other related domains**
- **A single European voice in international collaborations to influence global decisions and maintain open access to data**
- **Enhance European competitiveness in bioscience industries**
- **Address need for Increased Funding & its Coordination**

What might ELIXIR provide?

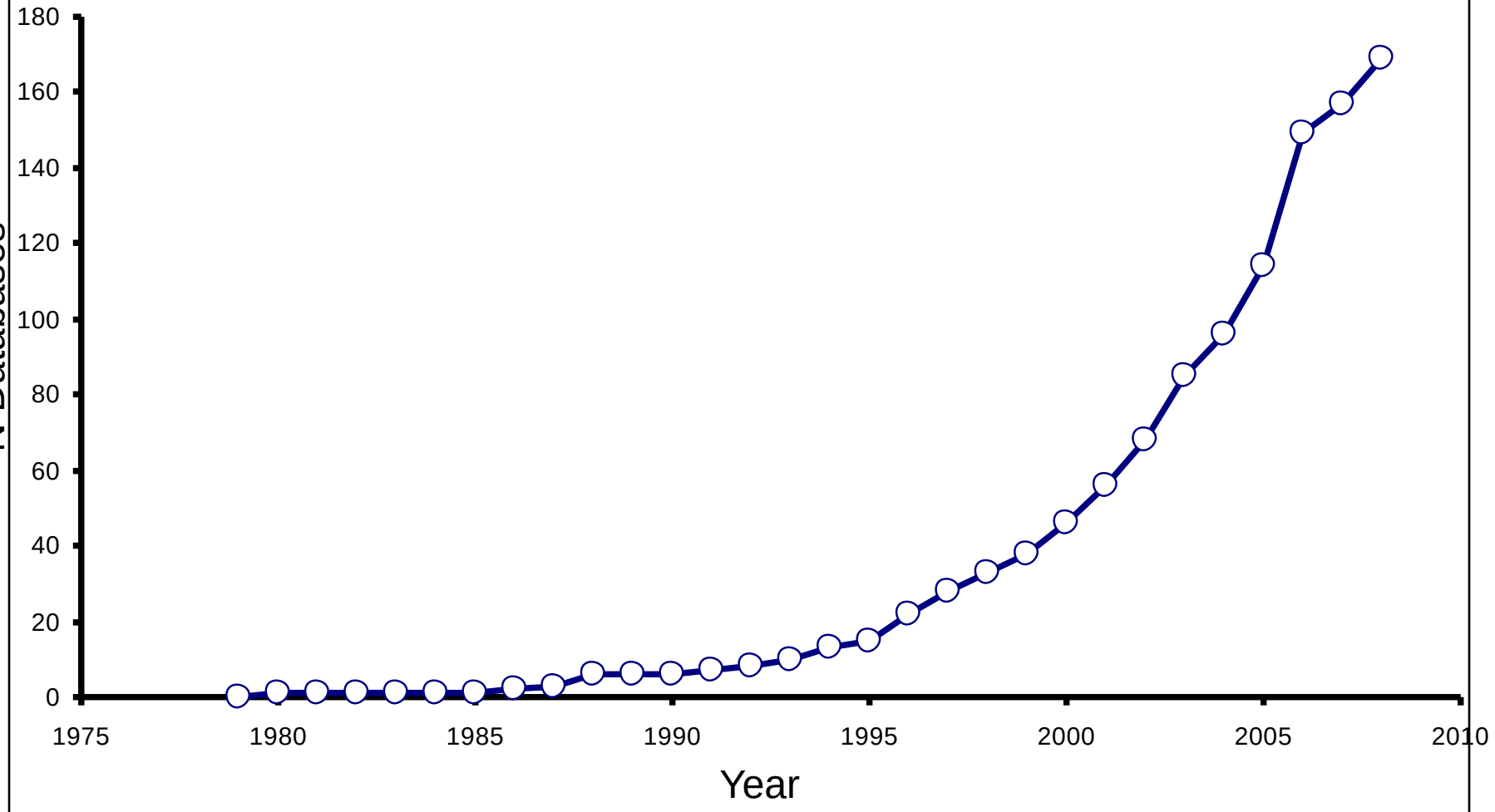
- Sustainable funding for key biological data resources
- A trans-national infrastructure for biological information and service providers, especially in new accession states
- A major upgrade of the current physical infrastructure, including construction of a European Biomolecular Data Centre.
- An infrastructure for tool integration

- The Preparatory Phase project has **two phases**:-
  - **Committee meetings of stakeholders to achieve consensus and make recommendations**
    - Jan 2008 – July 2009
    - Define scope and remit of ELIXIR
  - **Documentation and negotiation phase**
    - July 2009 – Dec 2010
    - Develop ‘Memorandum of Understanding’
    - Define funding and legal model

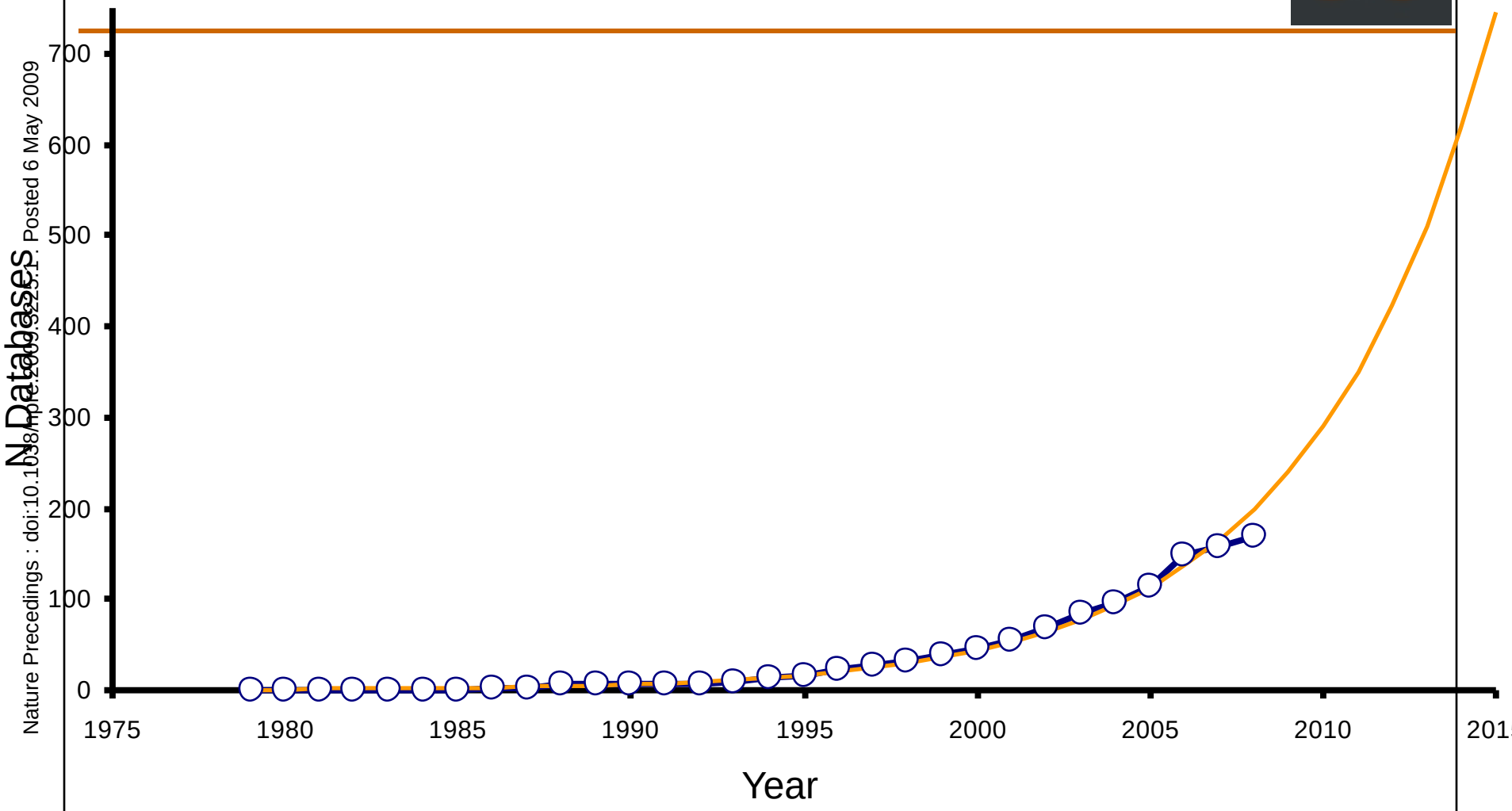
ELIXIR evolution



Age of Databases



2015?



ELIXIR: *a **sustainable** infrastructure for biological information in Europe*.

What might ELIXIR provide?

- An infrastructure for data curation?
  - Is this needed?
  - If so, what would this involve?
    - Could tools developed to curate one sort of data be useful in other resources?
    - Literature curation is common to almost all resources – how can this be improved?
    - How can links be automatically generated & updated
- You are the people to take this forward!

Overview

- Introduction: Personal experience with the PDB
- Curation @ EBI
- Future of Biological Information in Europe – ELIXIR
- Challenges for Curation

Challenges for Curation

- Flood of Sequence data
- Curating Biological Function
- Human Variation Data
- Metagenomic Data eg Microbiome
- Linking between Heterogenous Data Resources
- Universal ‘Sample’ Descriptions
- Distributed Data Resources? – Not if, but when!
- Improving links to the literature
- Better tools
  - For submission
  - For searching
- Community Annotation
  - The semantic web