Elixir

Data exchange between pan-European Scientific Databases

Open e-IRG Workshop in Prague

May 14-15, 2009

Andrew Lyall

Version 0.1

www.elixir-europe.org

EMBL-EBI

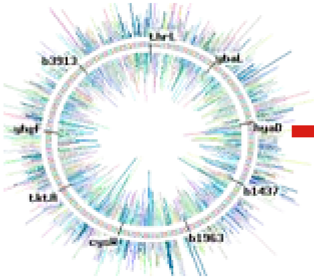
CAPACITIES

ESFRI

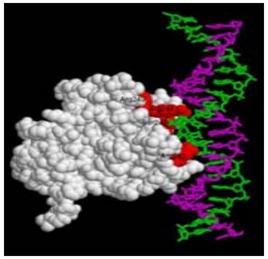
What is Elixir?

- An EU Framework 7 Preparatory Phase Project
- Coordinated by Prof Janet Thornton, Director EMBL-EBI
- To construct a plan for the operation of a ***sustainable*** infrastructure for biological information in Europe
- €4.5 million grant awarded May 2007, three year term
- 32 member consortium engaging many of Europe’s main bioinformatics funding agencies and research institutes
- Deliverables are memoranda of understanding to fund the implementation phase which could cost €500 million
- ***Requirement to distribute very large amounts of data around Europe***
- Interested parties should register as stake-holders via the ELIXIR Website: www.elixir-europe.org

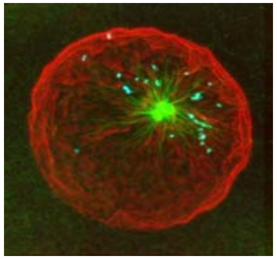
Modern biology requires integration.



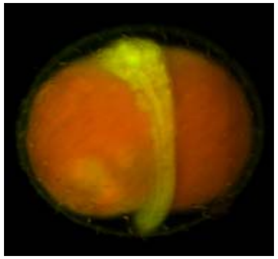
Genome



Protein



Cell



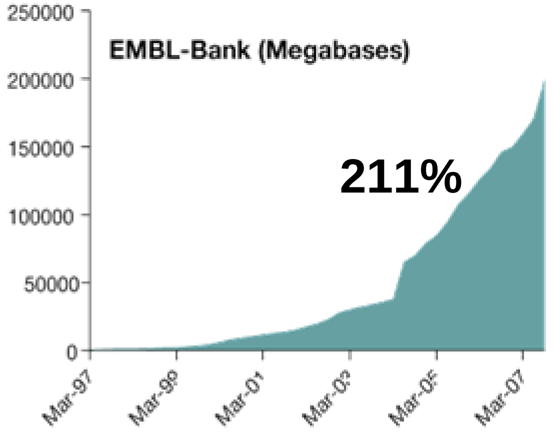
Embryo

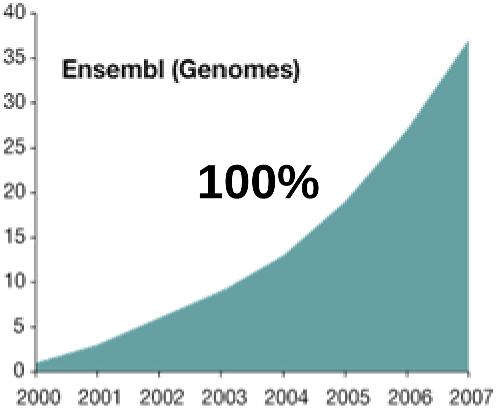
Fruitfly

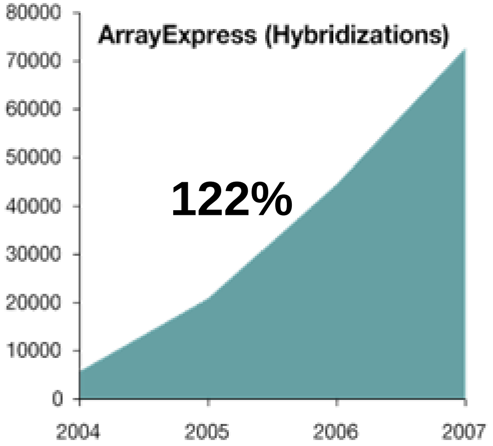
Mouse

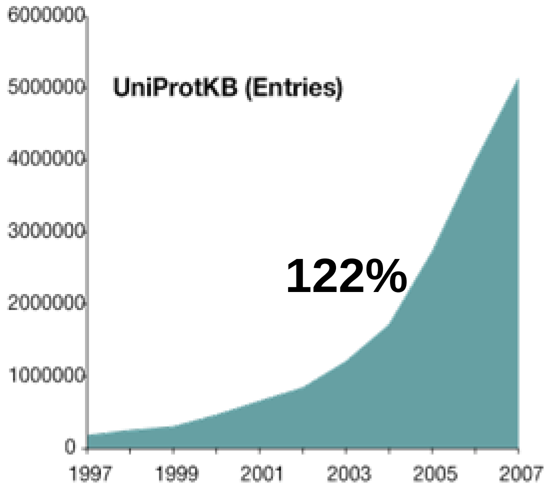
Development,
Ageing, Disease

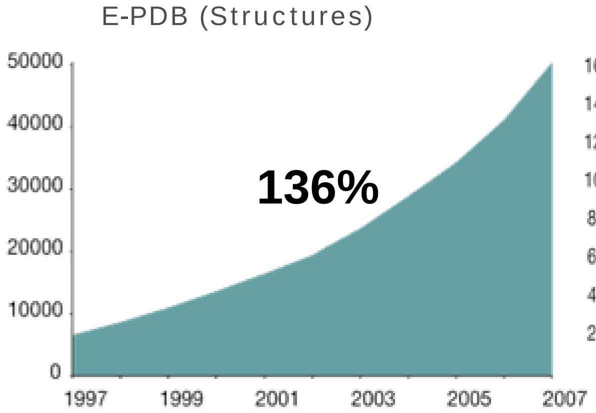
Database growth (2007/2006 %)

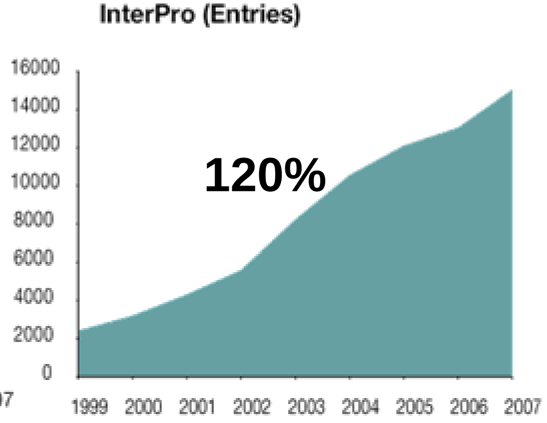




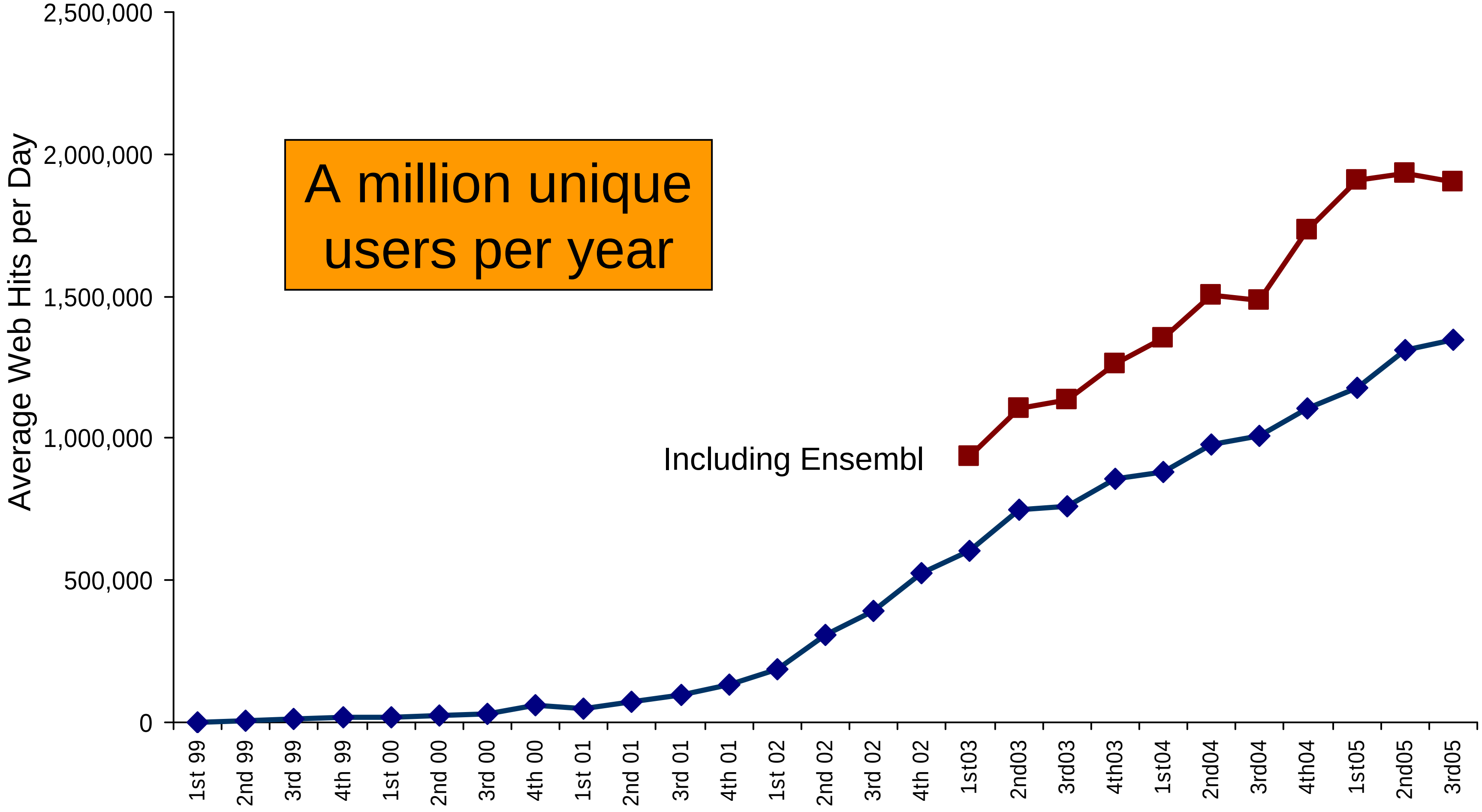






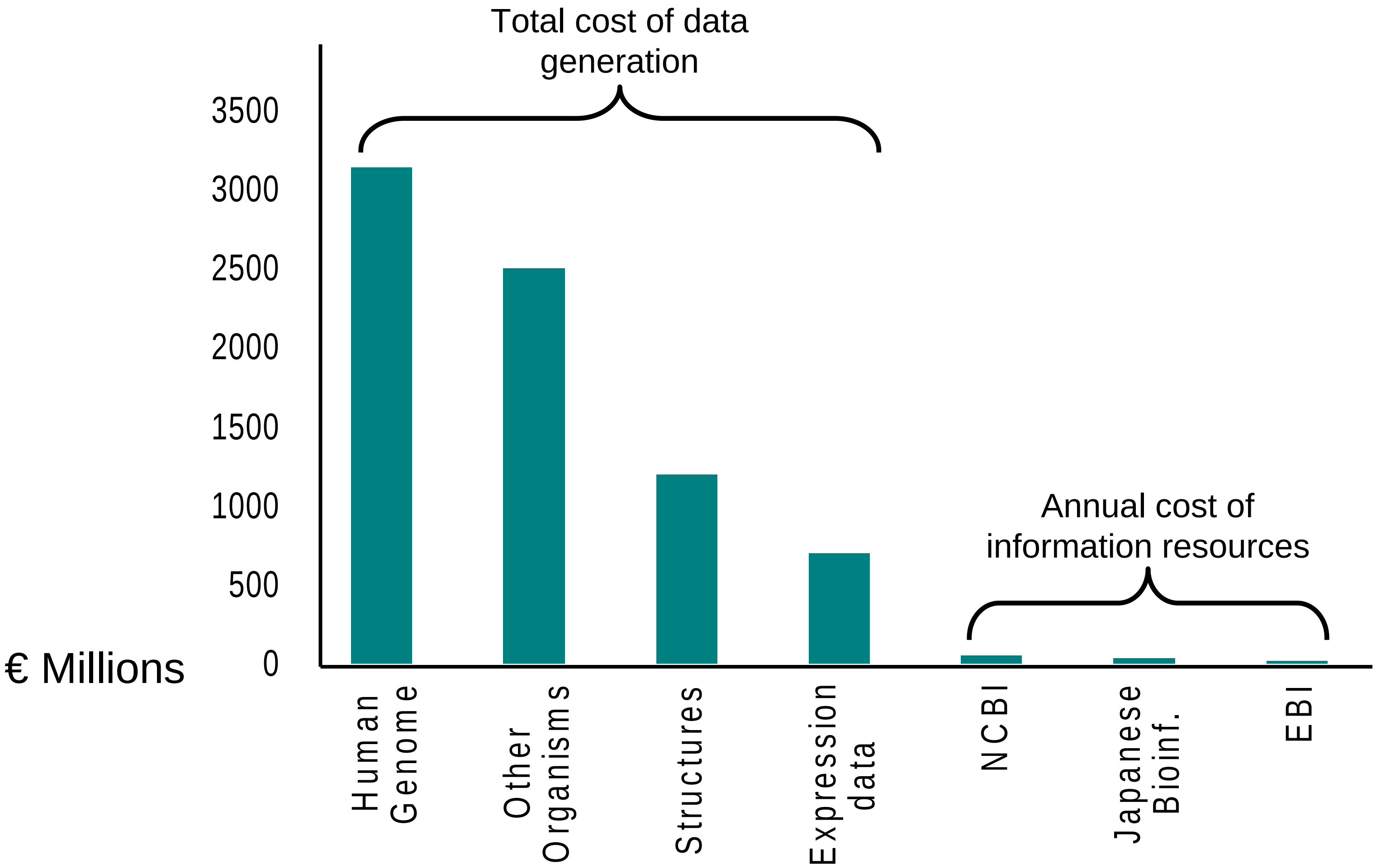


Very large user community



A million unique users per year

Good value for money

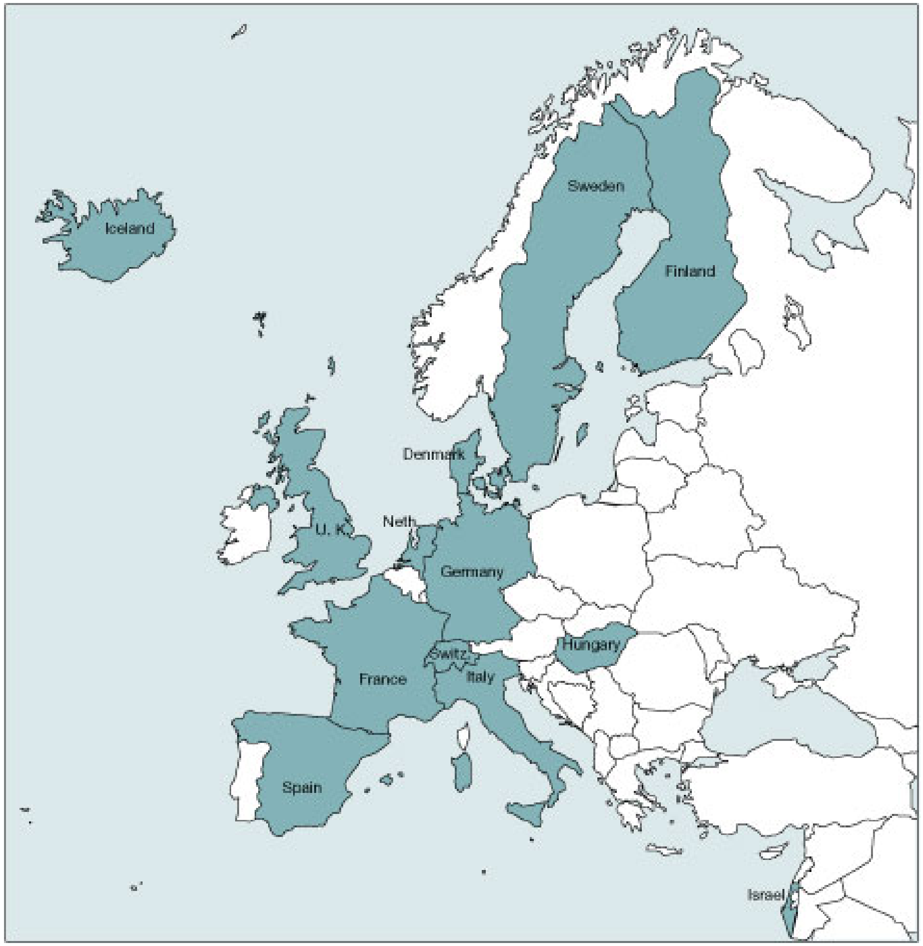


Elixir rationale

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

Members of the ELIXIR consortium

- There are 32 partners from 13 member states and associated countries
- 16 of the partners are funding agencies or Government Bodies
- 16 of the partners are scientific organisations or institutes
- There are expressions of interest from many others



Participants & Contacts 1

	Participant organisation name (& point of contact)	Short name	CC
1	EMBL - European Bioinformatics Institute (Prof. Janet Thornton, Dr. Dominic Clark).	EMBL-EBI	INO
2	Biotechnology and Biological Sciences Research Council (Dr. Alf Game)	BBSRC	UK
3	Federal Ministry of Education & Research (Dr. Elmar Nimmesgern)	BMBF	DE
4	Barcelona Supercomputing Center – Centro National de Supercomputacion (Prof. Modesto Orozco)	BSC	ES
5	Spanish National Cancer Research Centre (Prof. Alfonso Valencia)	CNIO	ES
6	Council for National Research (Dr. Giuseppe Martini)	CNR	IT
7	Center for Advanced Studies, Research and Development in Sardinia (Prof. Anna Tramontano)	CRS4	IT
8	CSC – Scientific Computing Ltd, Finnish Supercomputing Centre (Dr. Tommi Nyrönen)	CSC	FI
9	German Research Foundation (Dr. Nikolai Raffler)	DFG	DE
10	Danish Technical University (Prof. Søren Brunak)	DTU	DK
11	Erasmus Medical Centre (Prof. Johan van der Lei)	EMC	NL
12	Institute of Enzymology (Prof. Laszlo Patthy)	ENZIM	HU
13	Genome Espana (Dr José Luis Jorcano)	GE	ES
14	Forschungszentrum Fuer Umwelt und Gesunndheit GmbH (Prof. Hans-Werner Mewes)	GSF	DE
15	National Institute for Research in Computer Science & Control (Hugues Leroy)	INRIA	FR
16	Linköping University (Prof. Bengt Persson)	LiU	SE

Participants & Contacts 2

	Participant organisation name (& point of contact)	**Short name**	**CC**
17	Ministry Of Science & Technology (Dr. Mahmoud Taya)	MOST	IL
18	Medical Research Council (Dr. Mark Palmer)	MRC	UK
19	Natural Environment Research Council (Dr. Sarah Collinge)	NERC	UK
20	Netherlands Organisation for Scientific Research (Dr. Crétien Herben)	NWO	NL
21	The Icelandic Centre for Research (Dr. Rebekka Valsdóttir)	RANNIS	IC
22	Radboud University Nijmegen (Prof. Gert Vriend)	RU	NL
23	Wellcome Trust Sanger Institute – (Dr. Tim Hubbard)	SANGER	UK
24	Sardegna Ricerche (Dr. Luca Contini)	Sardegna Ricerche	IT
25	Swiss Institute of Bioinformatics (Prof. Amos Bairoch)	SIB	CH
26	Syngenta Ltd (Dr. Mark Forster)	Syngenta	UK
27	Technical University of Braunschweig (Prof. Dietmar Schomburg)	TU-BS	DE
28	University of Bordeaux 2 (Prof. Antoine de Daruvar)	UB2	FR
29	Swedish Research Council (Prof. Bengt Persson)	VR	SE
30	Wellcome Trust (Dr. Deborah Colson/Dr Alan Schafer)	Wellcome Trust	UK
31	Institut National de la Recherche Agronomique (Dr Christine Gaspin)	INRA	FR
32	Institut National de la Santé Et de la Recherché Médicale (Prof. Jean-Louis Coatrieux)	INSERM	FR

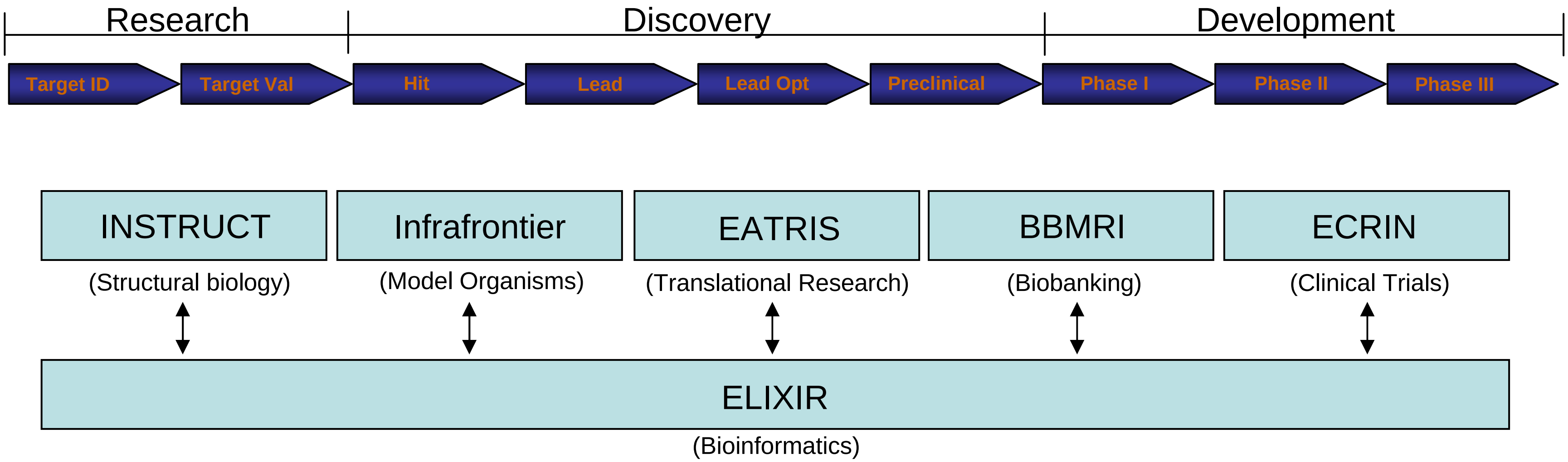
Is Elixir technically feasible?

Elixir does not depend for its success on any technology that has not been developed yet. However, it will be providing solutions to very demanding data management problems presented by things such as the 1000-genome project, the great increase in imaging of biological systems and the impending scale-up of structural and systems biology. We are thus conducting five technical feasibility studies that support the more challenging aspects of Elixir. More information on these studies is available from the Elixir Web Site.

1. Strategic Review of Cell Phenotype Image Data Resources.
2. Pilot of the use of European Supercomputing facilities for distributed processing of Bioinformatics data.
3. Assessment of European Resources for Systems Biology.
4. Search across heterogeneous distributed data resources (EB-eye).
5. Safe and ethical use of personal genetic information (European Genotype Archive).

ESFRI Biology RI proposals.

INSTRUCT	Integrated Structural Biology Infrastructure	300	25	2007	www.strubi.ox.ac.uk
Infrafrontier	Infrastructure for Phenomefrontier and Archivefrontier	320	36	2007	www.emma.rm.cnr.it
EATRIS	The European Advanced Translational Research Infrastructure	255	50	2010	http://www.eatris.eu/
BBMRI	European Biobanking And Biomolecular Resources	170	15	2009	www.biobanks.eu
ECRIN	Infrastructures For Clinical Trials And Biotherapy	36	5	2007	www.ecrin.org
ELIXIR	Upgrade Of European Bioinformatics Infrastructure	550	7	2007	www.ebi.ac.uk
		1631	**138**		



Research

Discovery

Development

Target ID

Target Val

Hit

Lead

Lead Opt

Preclinical

Phase I

Phase II

Phase III

INSTRUCT

(Structural biology)

Infrafrontier

(Model Organisms)

EATRIS

(Translational Research)

BBMRI

(Biobanking)

ECRIN

(Clinical Trials)

ELIXIR

(Bioinformatics)

What might Elixir be?

- A reliable ***distributed*** infrastructure to provide equality of access to biological information across all of Europe
- Sustainable funding for the ***core*** European biological data collections (genomes, sequences, structures etc)
- Sustainable funding for the ***global*** biological data collaborations (UniProt, ww-PDB, INSDC etc)
- Processes for

  developing ***new*** core data collections

  supporting ***interoperability*** of bioinformatics tools

  developing bioinformatics ***standards*** and ***ontologies***
- ***Enhanced*** use of biological information in Academic Research, the Pharmaceutical Industry, Biotechnology, Agriculture and for the Protection of the Environment

Attributes of core data collections

- Universally relevant to biology and medicine
- Journals insist on data deposition as a condition for publication
- Very, very large user communities
- Aim to be complete collections with Global significance
- Exchange with other data centres ensures completeness
- Science is stable enough to allow standardisation of data structures
- Host institute needs to be involved in standards development
- Support requires substantial institutional commitment

Core data collections at EMBL-EBI

1. **European-PDB** — the European partner in the wwPDB Macromolecular Structures Database.
2. **UniProt** — the world’s definitive collection of protein sequence data.
3. **EMBL-Bank** — the European instance of the global archive of nucleotide sequence data.
4. **Ensembl** — a world leader in the provision of annotated eukaryotic genomes.
5. **ArrayExpress** — a major public repository for microarray data.
6. **InterPro** — a database of protein families, domains and functional sites which aggregates such information from a large number of collaborators.

International Collaborations: wwPDB

- World-Wide Protein Data Bank
- Global archive of protein and other macromolecular structures
- In existence for nearly 40 years, currently a collaboration between
  1. EMLB-EBI: European-PDB (Molecular Structures Database)
  2. RCSB PDB: Protein Data Bank of the Research Collaboratory for Structural Bioinformatics (A consortium of US Universities)
  3. PDBj: Protein Data Bank Japan
  4. BMRB: Biological Magnetic Resonance Data Bank (University of Wisconsin-Madison)

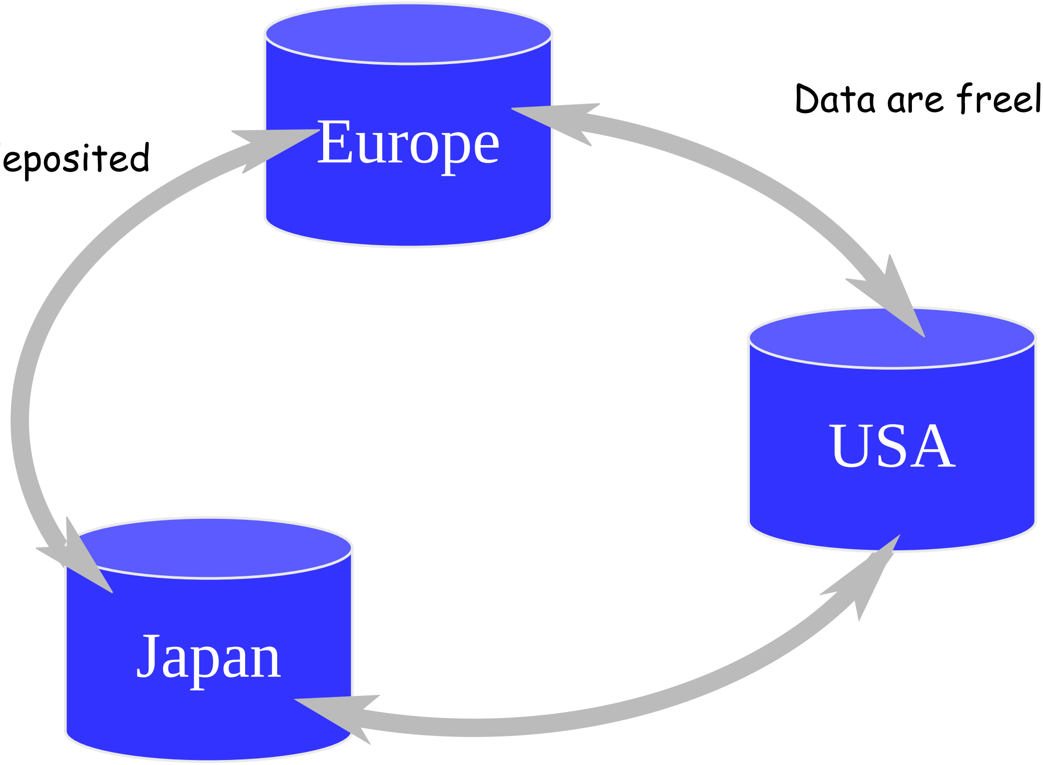
International Collaborations: UniProt

- UniProt
- A collaborations lasting many years between
  1. trEMBL: Translated EMBL at EMBL-EBI
  2. PIR: Protein Information Resource at Georgetown University Medical Centre
  3. SWISSPROT: Swiss Institute of Bioinformatics

International Collaborations: INSD

- **I**nternational **N**ucleotide **S**equence **D**atabase
- The worlds archive of Nucleic Acid Sequence Data
- A collaboration lasting nearly 20 years between:
  1. DDBJ: DNA Data Bank of Japan
  2. EMBL-Bank: Nucleotide sequence database of the EMBL-EBI
  3. NCBI GenBank: National Centre for Biotechnology Information USA

Global context



Data are freely deposited

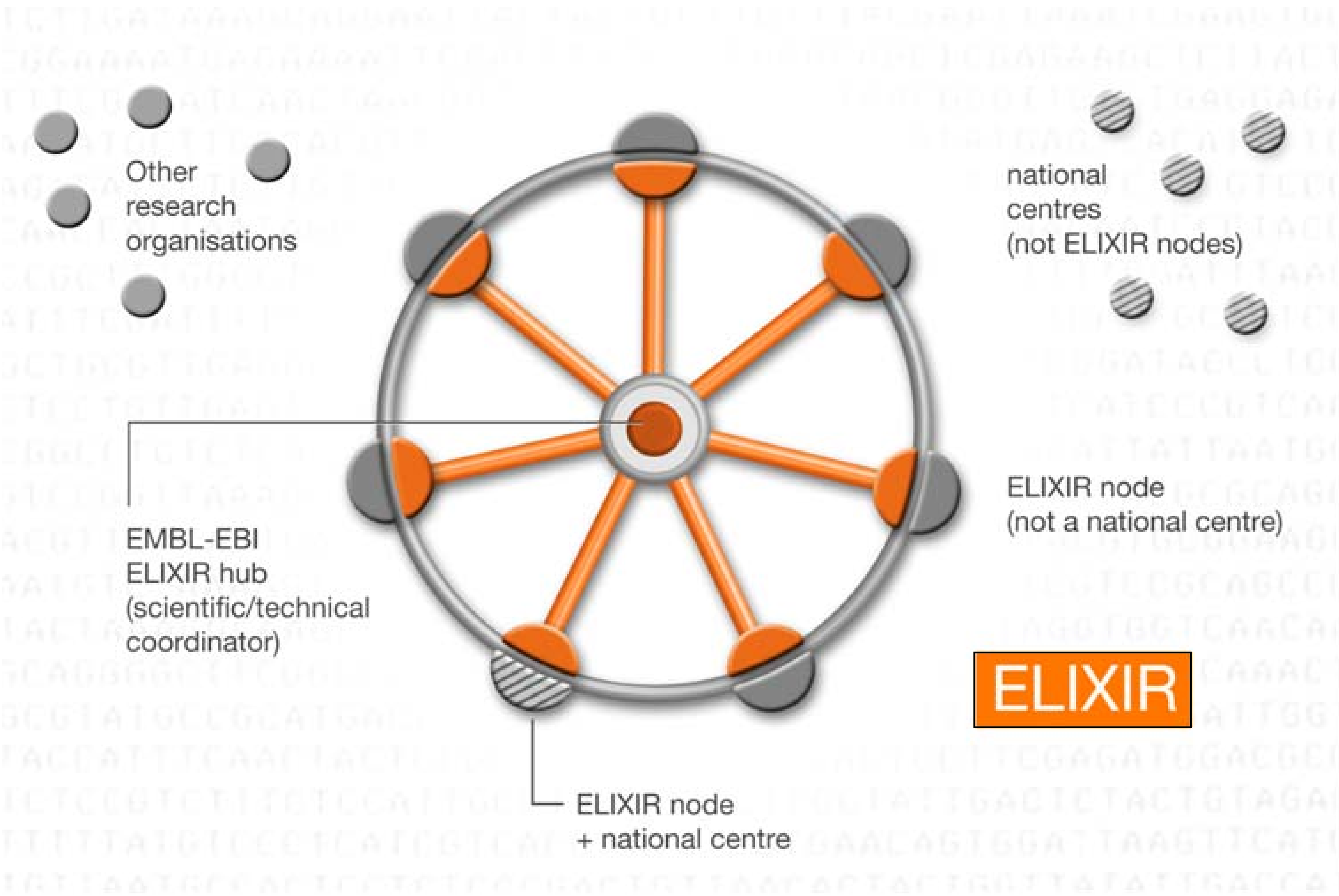
Data are freely exchanged daily

Data are made freely available to all

A Reliable Distributed Infrastructure

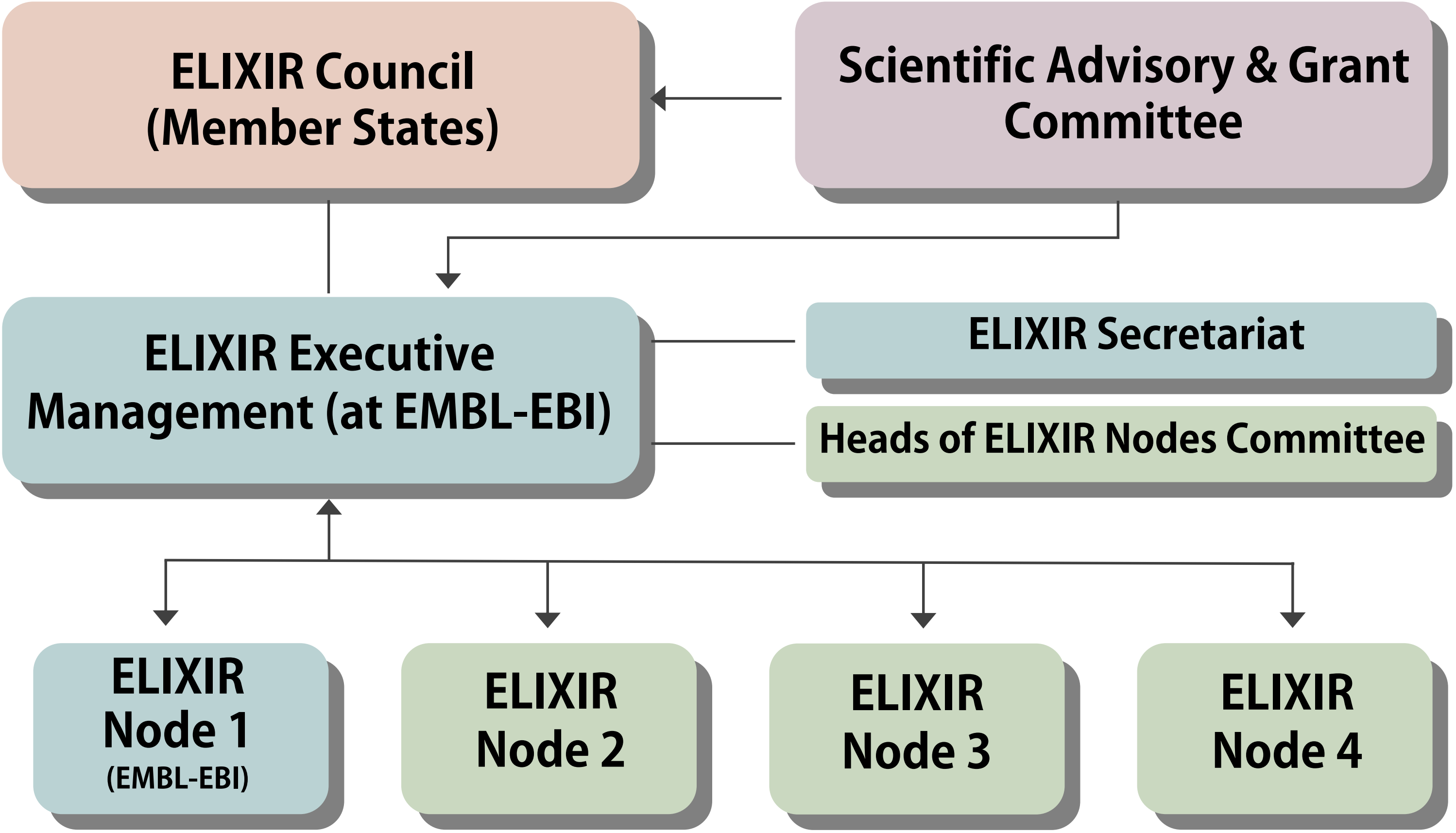
- Elixir will be constructed by enhancing and linking existing infrastructures in the member states.
- It will integrate member state infrastructures into a single infrastructure or a ‘Grid’.
- This will form a ‘Hub and Node’ structure
- Datasets will be assemble at the Hub
- Nodes may act as centres for collecting data
- Data will be distributed from the Hub to the Nodes

ELIXIR Technical Structure



Other research organisations
national centres (not ELIXIR nodes)
EMBL-EBI ELIXIR hub (scientific/technical coordinator)
ELIXIR node (not a national centre)
ELIXIR
ELIXIR node + national centre

ELIXIR Organisation


ELIXIR Council (Member States)

Scientific Advisory & Grant Committee

ELIXIR Executive Management (at EMBL-EBI)

ELIXIR Secretariat

Heads of ELIXIR Nodes Committee

ELIXIR Node 1 (EMBL-EBI)

ELIXIR Node 2

ELIXIR Node 3

ELIXIR Node 4

Properties of Nodes

An ELIXIR Node will be :-

- a legal entity or will be represented by a Legal Entity
- eligible for applying for and receiving funding
- able to join the ELIXIR Organisation
- capable of supporting one or more Components of the ELIXIR Infrastructure
- capable of entering into a contract with the ELIXIR Hub
- committed to delivering at least one component of ELIXIR

Components of the Infrastructure

A biomolecular data collection

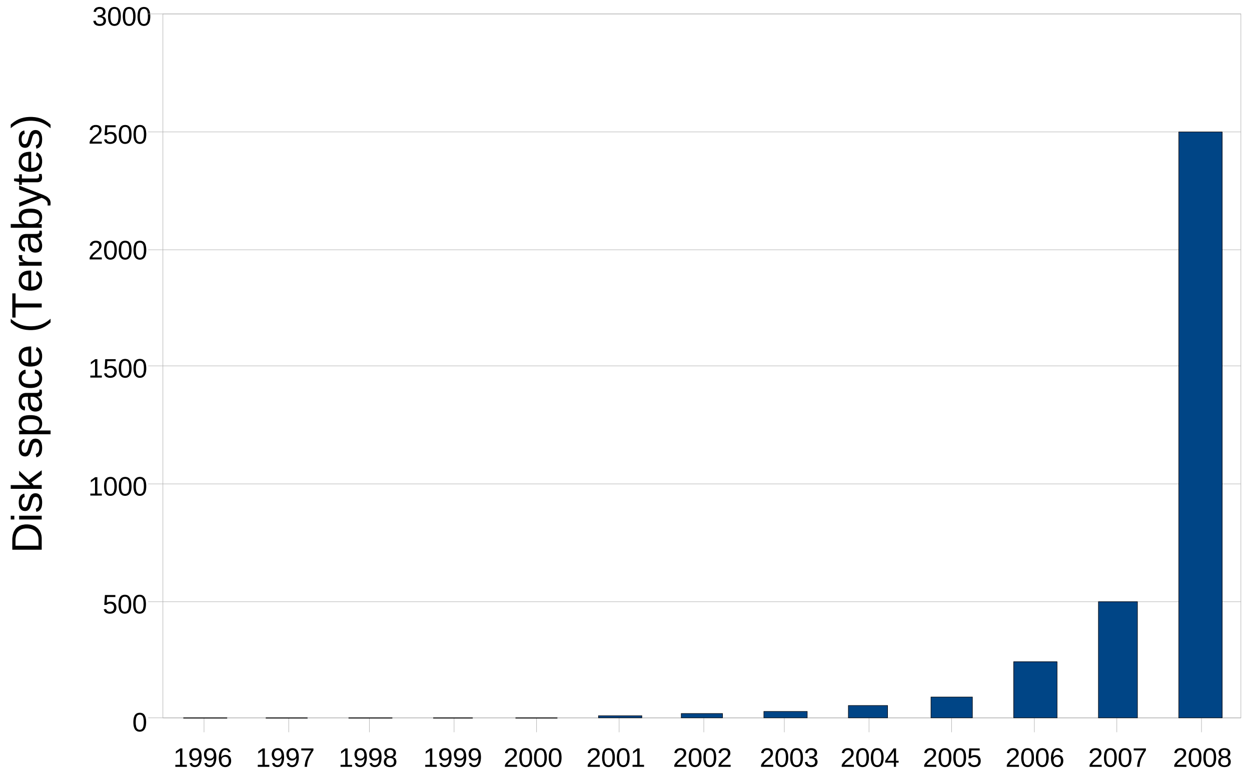
A means to deploy one or more biomolecular data collections

- A compute resource
- A storage resource

A means to access one or more biomolecular data collections

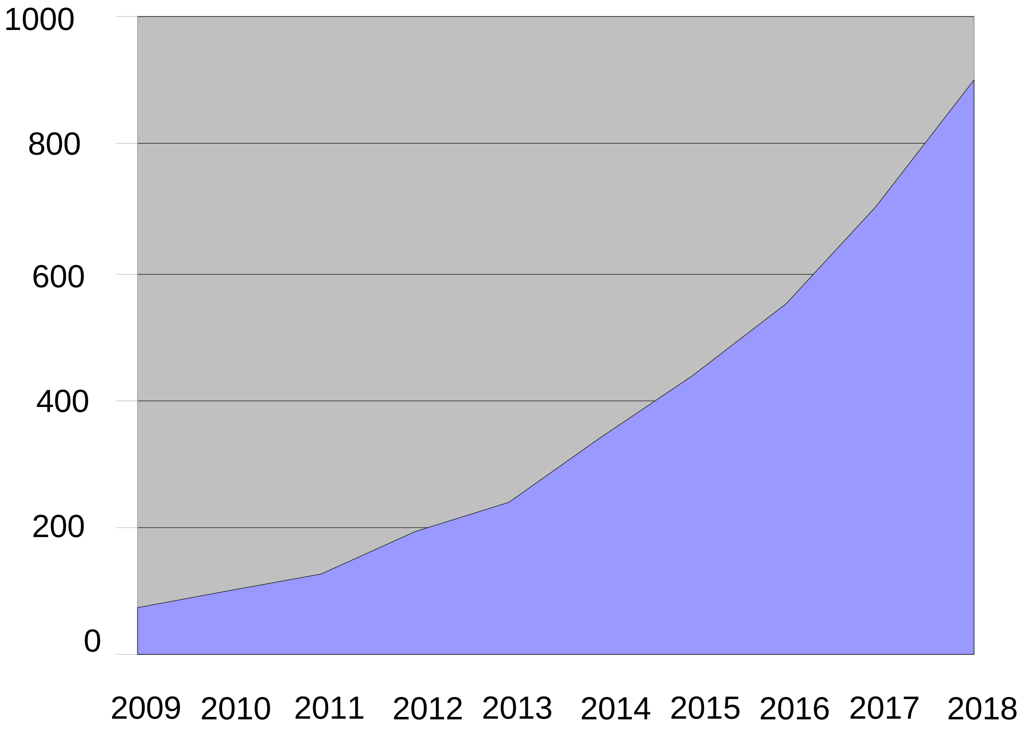
- Data access tools
- A catalogue or registry of data collections and tools
- A tools benchmarking capability
- A training/outreach resource
- A standards resource

Historical storage at EMBL-EBI



April 2009 Storage is 4.5 Petabytes!

Projected racks in EMBL-EBI data centre



Summary

- New data-generating technologies are transforming biology
- There will be many new data-generating projects
  - Thousand Genome Project
  - Cancer Genome Project
  - Etc. etc…
- Europe needs e-Infrastructure for biology as well as for physics
- Will be needed for meeting the European Grand Challenges
  - Healthcare for an aging population
  - Sustainable food supply
  - Environmental protection
  - Competitive Pharmaceutical and Lifescience Industries

The 1000 Genome Project

- A deep catalogue of human variation to provide a better baseline to underpin human genetics
- Specific Goals
  - Discover genetic variation in major human populations to support research in clinical and population genetics
  - Develop analysis tools to support ultra-high-throughput sequence generation
  - Understand the evolutionary pressures on the human genome
- Data Generation
  - 9 sequencing centers representing 4 countries (UK, USA, China, Germany)
  - Currently 3 sequencing technologies being used
- Dataset is of interest to scientists and clinicians across all of Europe
- ***Data set size – 500 terabytes***
- http://www.1000genomes.org/