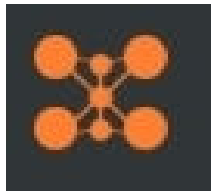
ELIXIR

ESFRI & e-Infrastructure Collaborations

Presented at EGEE'09 Barcelona
October 2009
Andrew Lyall
Version 0.1

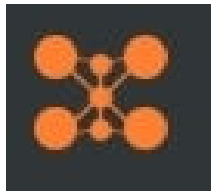
www.elixir-europe.org

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
- Interested parties should register as stake-holders via the ELIXIR Website: www.elixir-europe.org

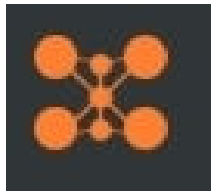
ELIXIR Work Packages.



Elixir is organised into 14 work packages which have committees of (mainly) European experts associated with them. It is organising two surveys, one of users and one of data-providers, and five technical-feasibility studies. The Elixir Steering Committee is associated with WP1 and has oversight of the whole project. WP3 has four committees; for Bioinformatics Communities, for Data Providers, for Industry and for Interactions with the rest of the World (International). There will be regular Stakeholder meetings intended to encourage the widest possible participation.

1. Project management
2. Data providers
3. User communities
4. Organisation and Legal
5. Funding
6. Physical infrastructure
7. Data interoperability
8. Literature
9. Healthcare
10. Chemistry & Environment
11. Training
12. Tools integration
13. Feasibility studies
14. Reporting and negotiation

Summary of consultation phase

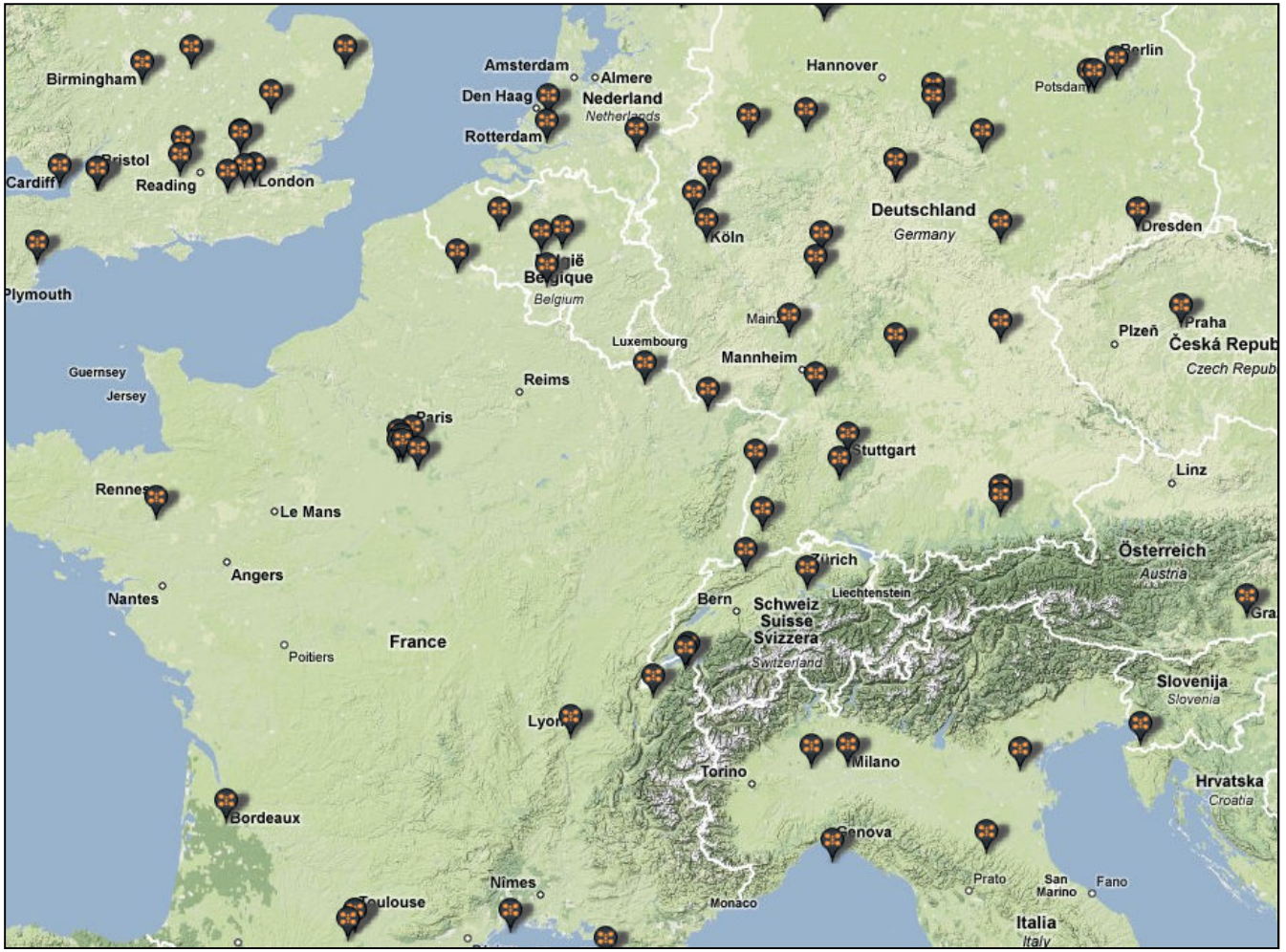


- Three Stakeholders Meetings
- Many Steering Committee Meetings
- Data providers survey
- Bioinformatics users survey
- Reports from Work Packages and Feasibility Studies.
- Numerous visits
- ELIXIR placed on Member State Infrastructure Roadmaps
- Sweden first country to commit
- UK commit £10M pulse of capital from LFCF Roadmap
- Reports, results of surveys, details of visits, etc. are on the web site.

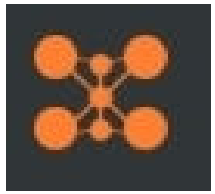
http://www.elixir-europe.org/

Visits during consultation phase.

Sites of ELIXIR survey data providers



Sweden is first country to commit



Sweden to be the first country to pledge long-term funding for ELIXIR

Posted on Tue Jun 23, 2009 7:32 am

The Swedish funding agencies (the Swedish Energy Agency, Fas, Formas, the Swedish Research Council and VINNOVA) have suggested that the Government allocate a total of **19 million SEK** (1.7 million Euro) over three years to ELIXIR & the Swedish Bioinformatics Infrastructure for Life Sciences (BILS), which would make Sweden the first country to secure long-term funding for ELIXIR. A final decision along these lines is expected during the Autumn 2009.

Contact: Prof. Bengt Persson, Linköping University & Karolinska Institutet

Web site: www.bils.se

UK commits to ELIXIR

UK leads European research programme with £10M investment in bioscience data handling capacity

Posted on Tue Aug 25, 2009 8:35 am

The UK has made its first substantial commitment to ELIXIR with a £10M investment by the Biotechnology and Biological Sciences Research Council (BBSRC). BBSRC has awarded funding to the European Molecular Biology Laboratory's European Bioinformatics Institute to permit a dramatic increase in the institute's data storage and handling capacity. The funding is the first step in developing the existing data resources and IT infrastructure of EMBL-EBI towards its planned role as the central hub of the emerging European Life-Science Infrastructure for Biological Information.

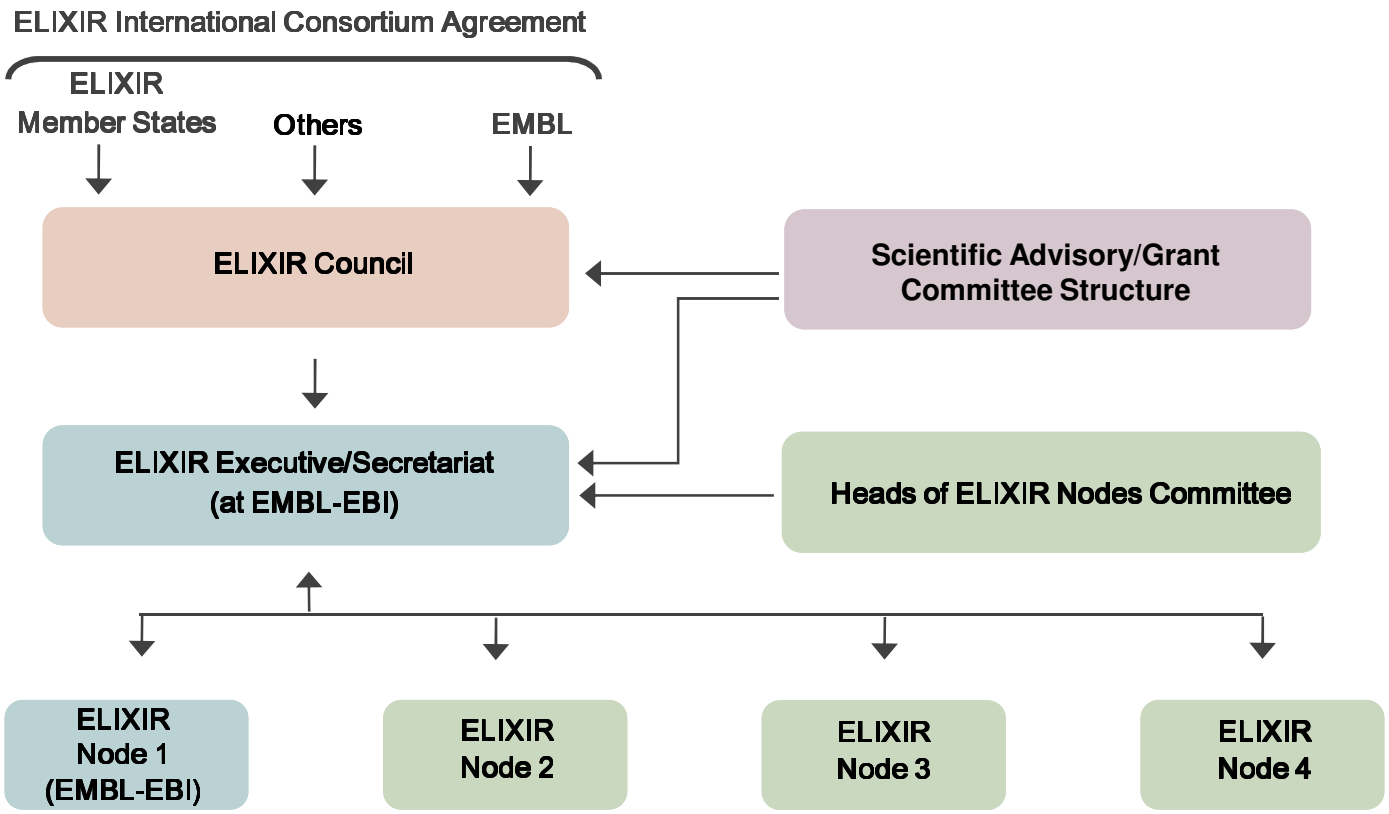
Contact: Matt Goode, BBSRC External Relations

Web site: www.bbsrc.ac.uk

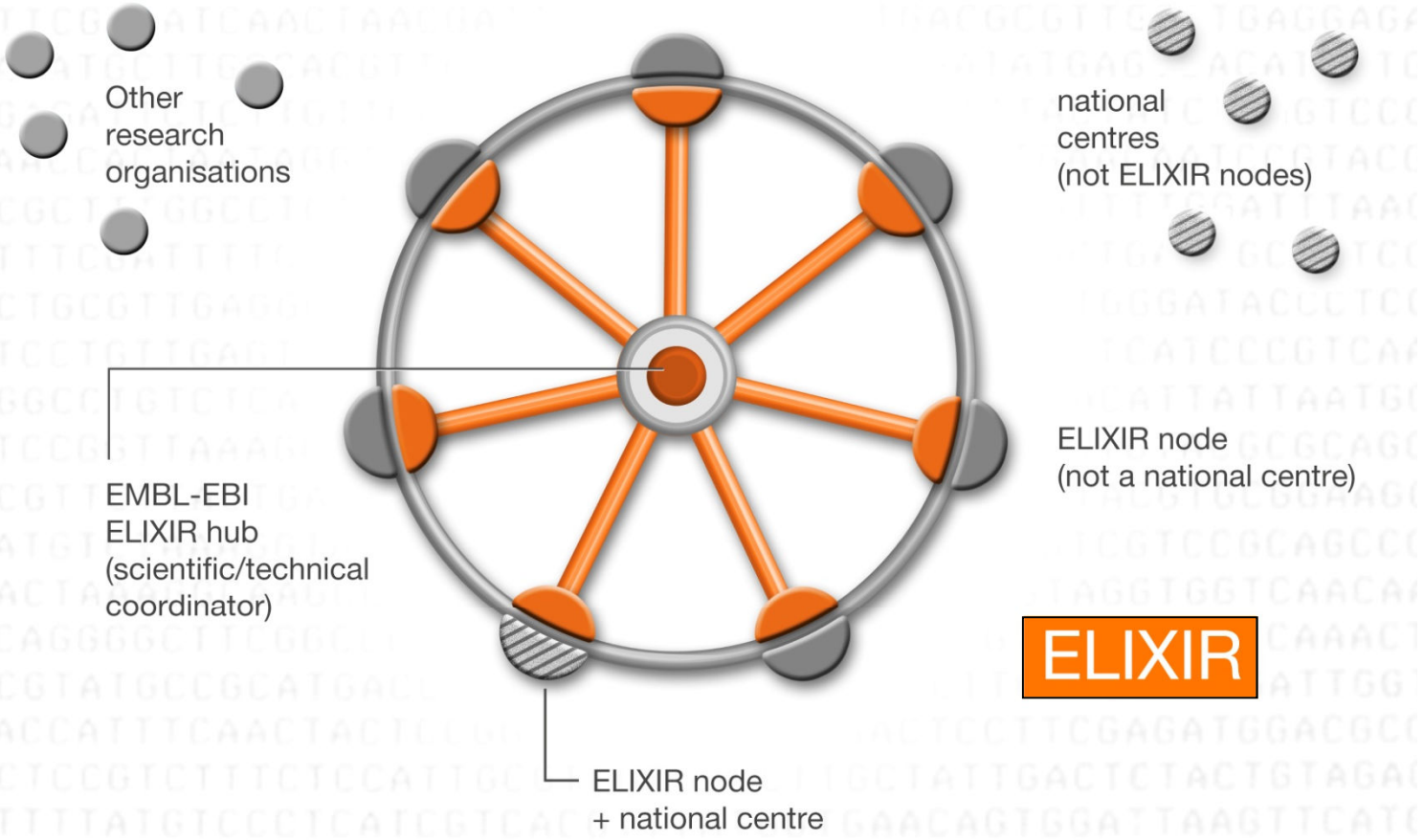
ELIXIR Legal Personality

- Initially ELIXIR is most likely to be an EMBL “Special Project” although the decision has not been taken yet.
- In due course this will probably be transferred to an “ERIC”
- This approach will allow a quick start for the construction phase.
- Early adoption of “ERIC” was considered high risk.
- Decision to change will be taken by ELIXIR Management.
- We have taken legal advise on this.
- Bearing in mind the critical importance of ELIXIR for Europe, this is considered the safest way to proceed.

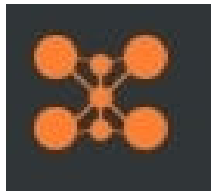
ELIXIR Management structure



ELIXIR Scientific & Technical Structure

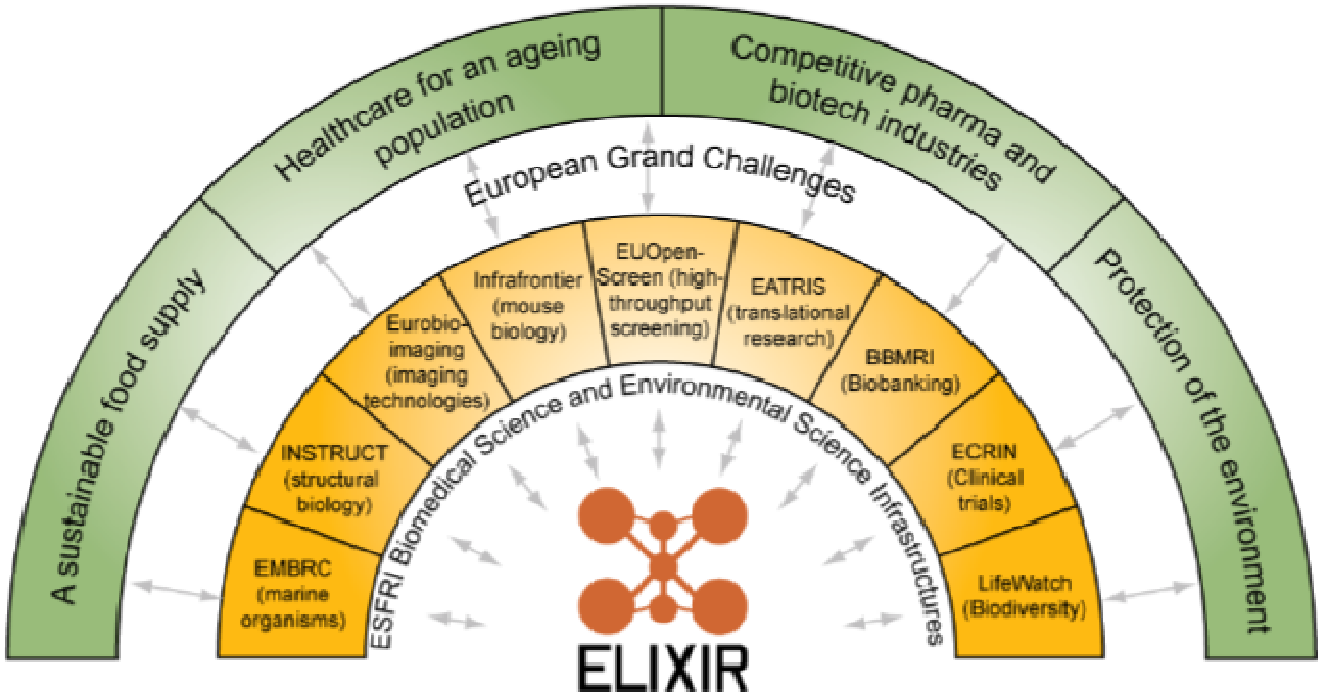


European Grand Challenges



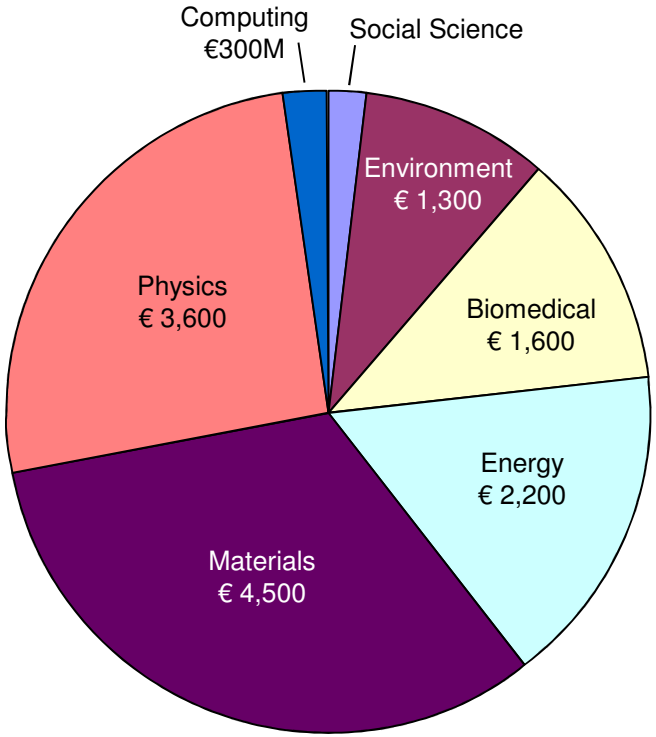
- Healthcare for an aging population
- A sustainable food supply
- An internationally competitive life-sciences industrial sector
- Protection of the environment.
- A sustainable energy supply

ELIXIR Support of the European Grand Challenges



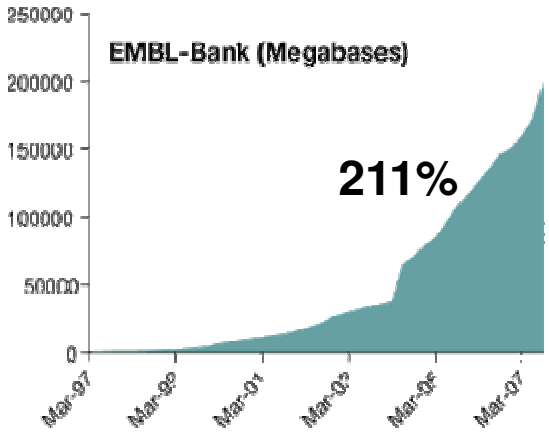
ELIXIR supports the European Grand Challenges by providing Infrastructure for the other ESFRI Biology Projects.

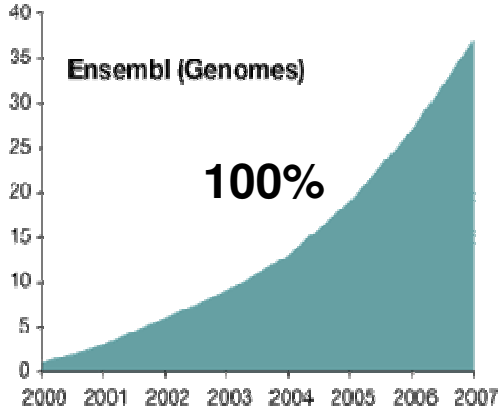
Cost of 35 Mature ESFRI RI Projects

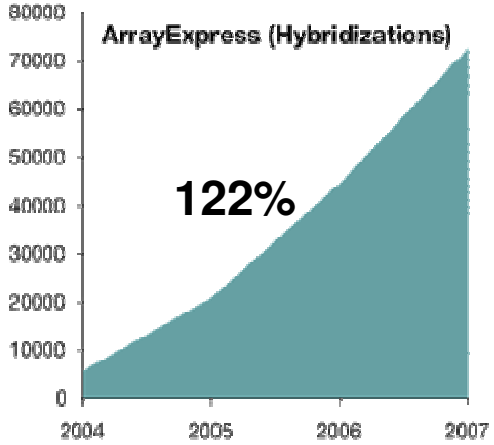


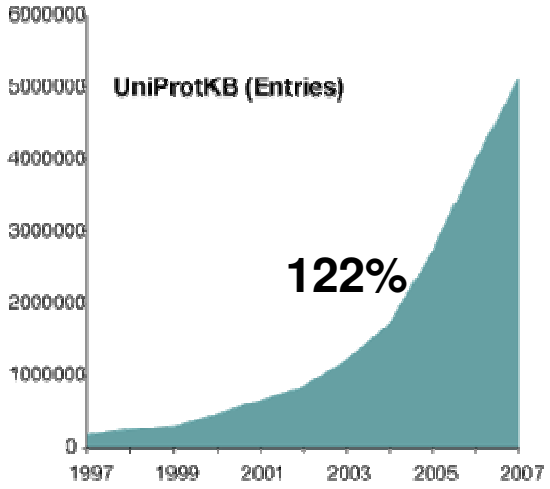
Total Capital Cost = €13,696 Million

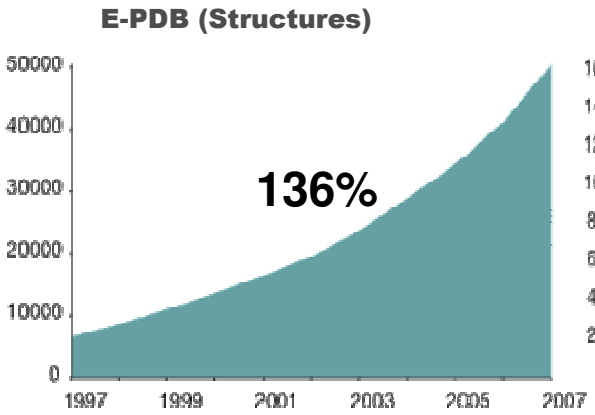
Database growth (2007/2006 %)

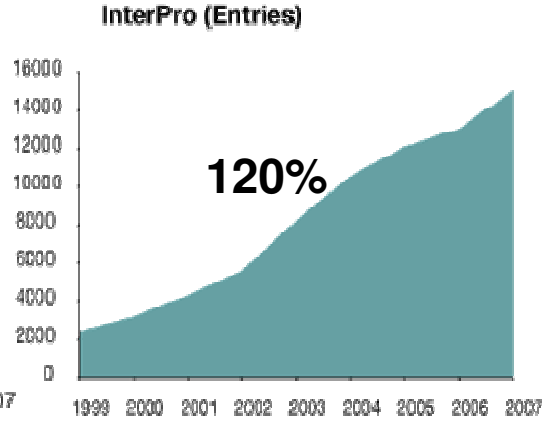




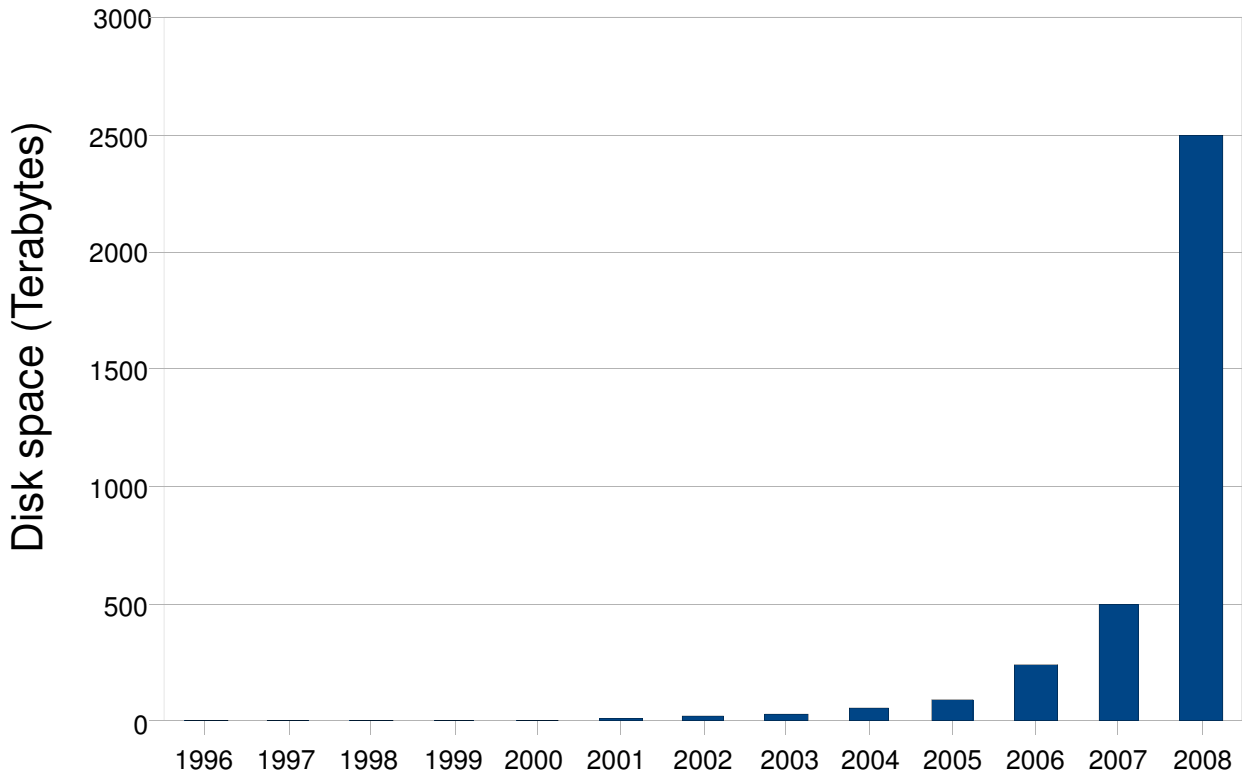






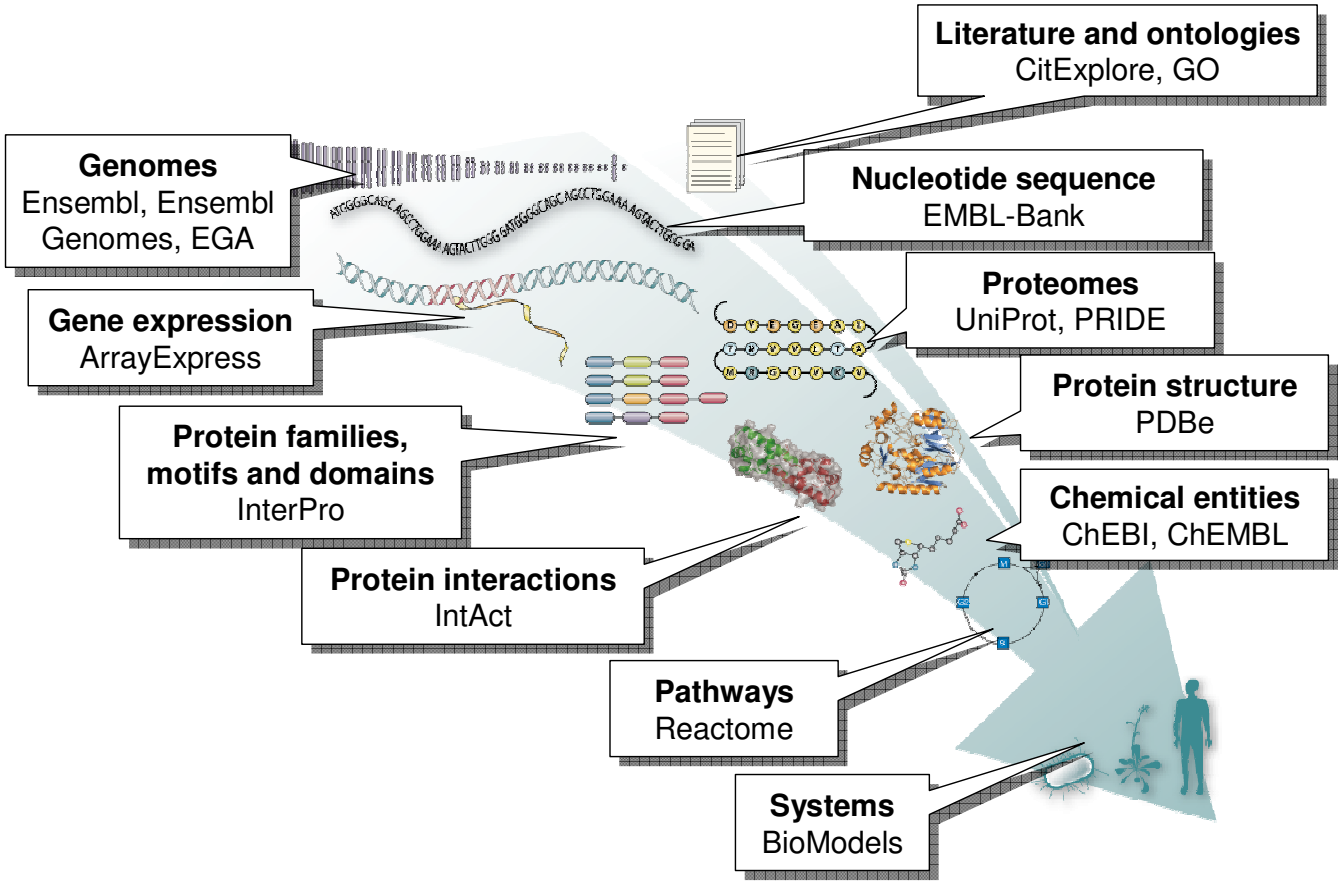


Historical storage at EMBL-EBI

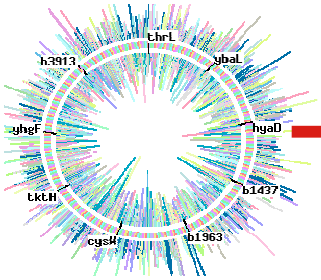


* September 2009 Storage is has reached ***4.5 Petabytes!***

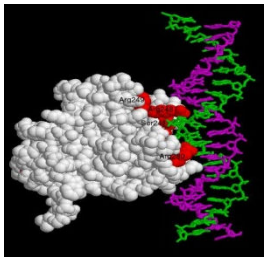
Not just DNA sequence data.



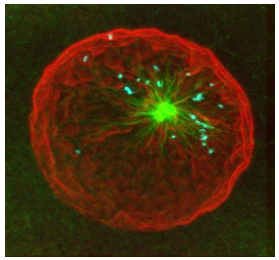
Modern biology requires data integration too!



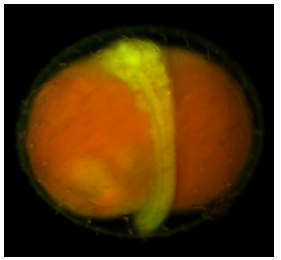
Genome



Protein



Cell



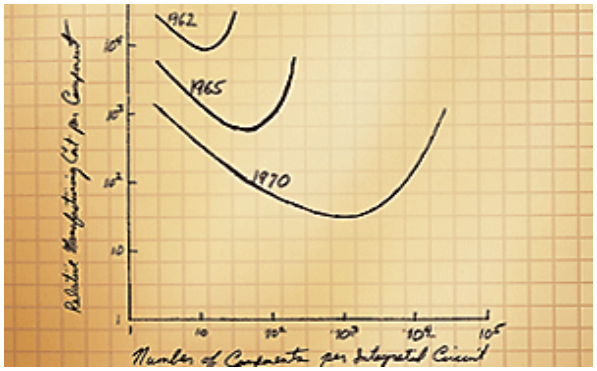
Embryo

Fruitfly

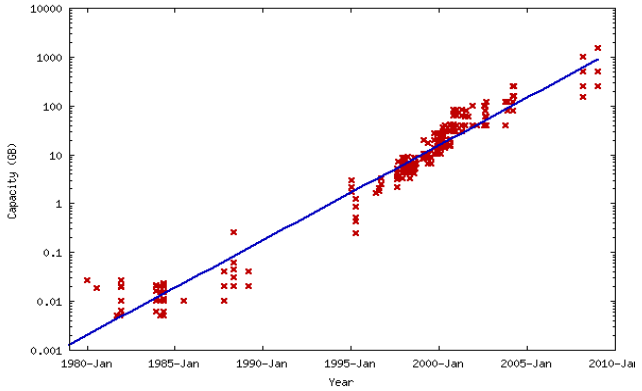
Mouse

Development,
Ageing, Disease

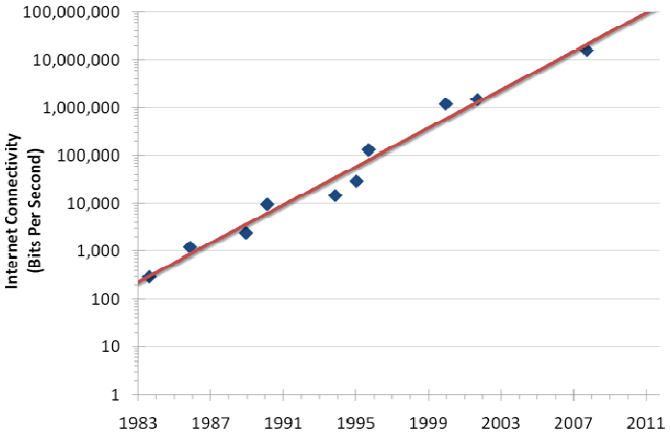
Data growth exceeds growth in IT capability



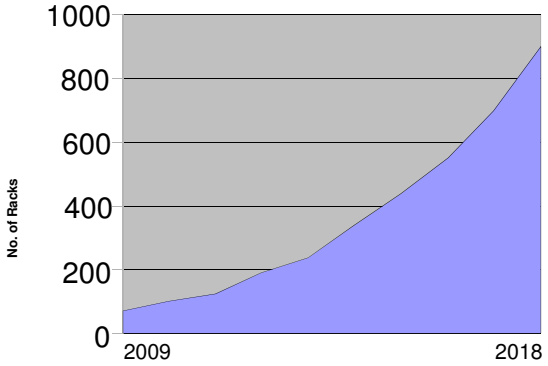
CPU Power doubles <> 24 months (Moores Law)



Disk capacity doubles <> 18 months

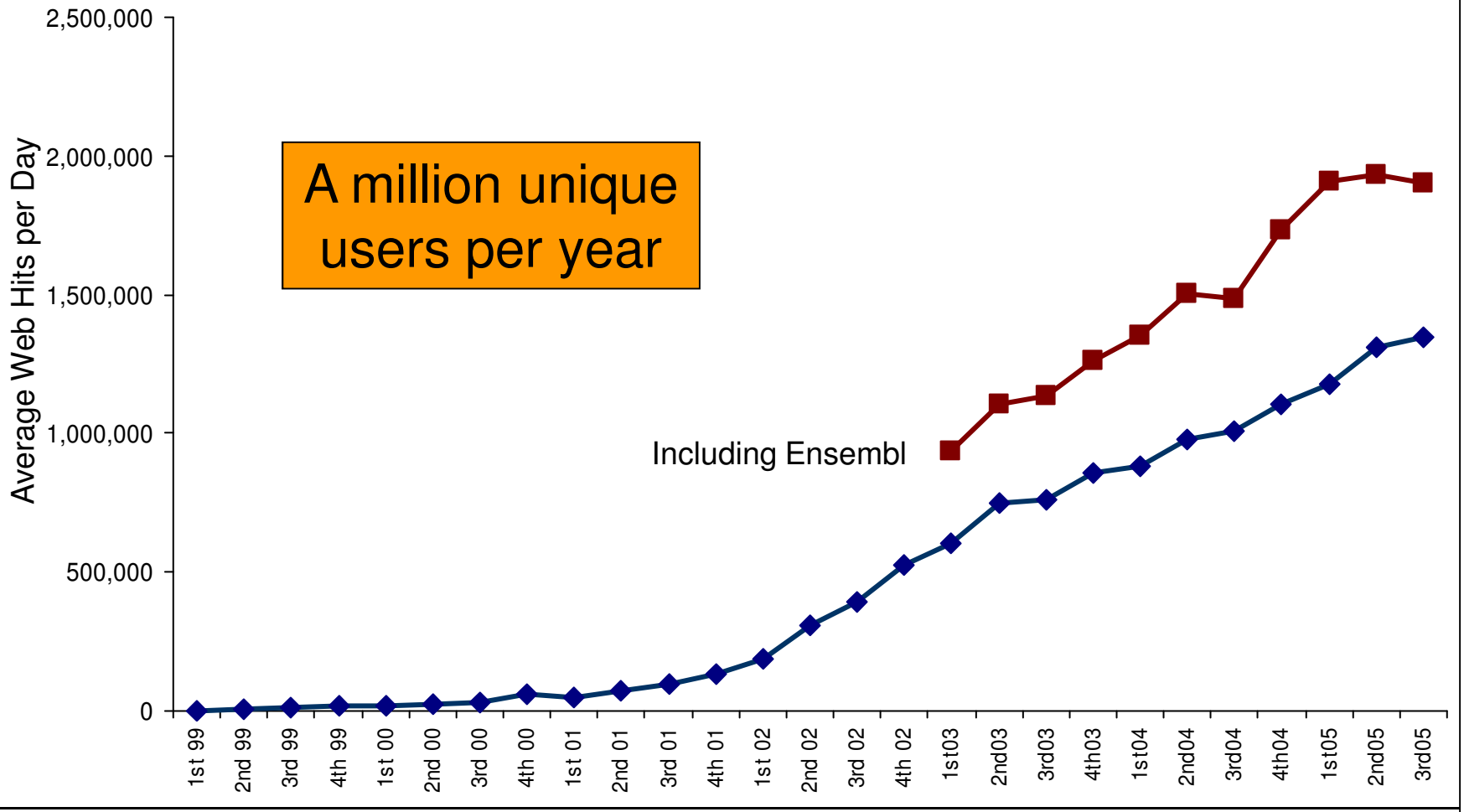


Network bandwidth doubles <> 20 months

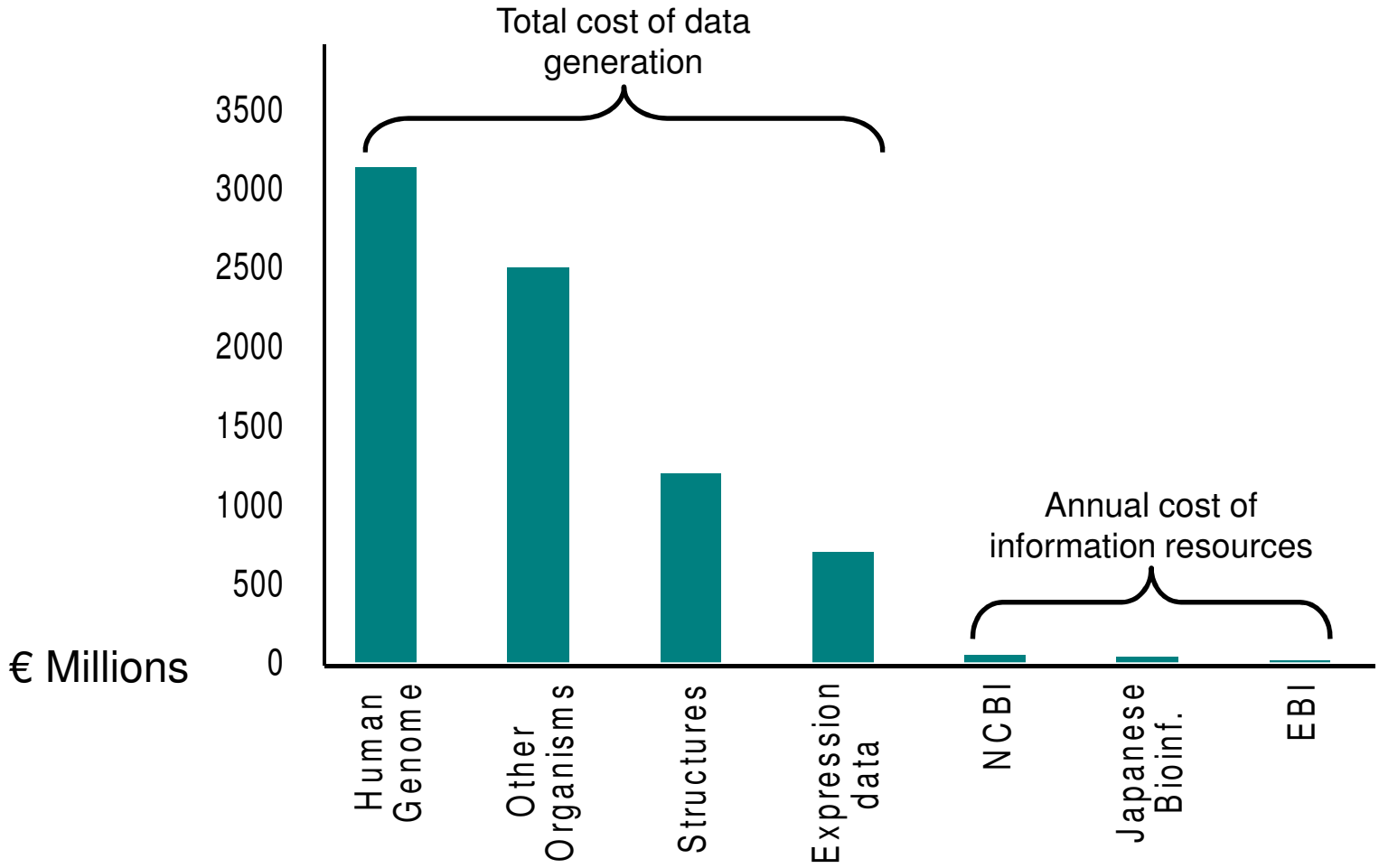


Racks in EBI machine room double <> 12 months

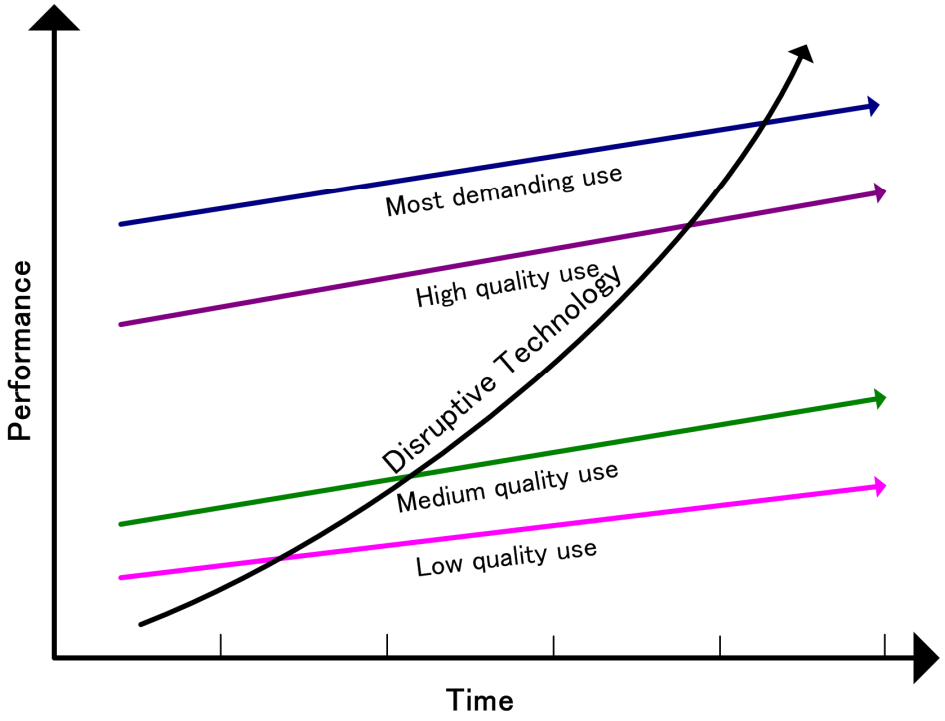
Very large user community



Good value for money



Disruptive technologies.



"A technology becomes disruptive when the rate at which it improves exceeds the rate at which users can adapt to the new performance."

The Innovator's Dilemma. Clayton M. Christensen. Harvard Press. 1997

Disruptive technologies in biology

- Next-generation DNA sequencing
  - Data will be 1,000 <> 1,000,000 times cheaper to produce
  - Data production rates will be 1,000 <> 1,000,000 more by the end of the ESFRI period.
- Protein sequencing by Mass Spectrometry may also be disruptive
- There will probably be others
  - Macromolecular structure determination by Electron Microscopy
  - Imaging of various kinds
  - etc

Exponential growth in data…

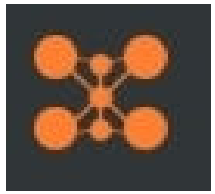
Cannot equate to exponential growth in funding, so

1. Link budgets for data generation and data processing
   - Only produce as much data as you can deal with

2. Take steps to control staff growth
   - Automation of annotation and curation
   - Implement distributed annotation (DAS)
   - Use web services and distributed resources
   - Support for metadata deposition

3. Take steps to control IT resource requirements.
   - Develop policies for which data are to be kept (& which not)
   - Develop data compression techniques

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).

Pilot use of European Supercomputing Facilities

Partners:

- Sarah Hunter, Antony Quinn

EMBL

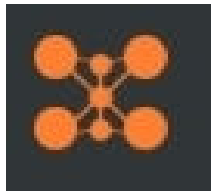
- Modesto Orozco, Josep Luís Gelpí, David Torrents, Romina Royo, Enric Tejedor

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

- Tommi Nyronen, Marko Myllynen, Pekka Savola, Jarno Tuimala, Olli Tourunen, Samuli Saarinen

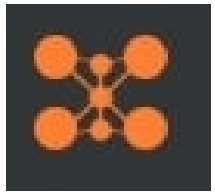
CSC

Results of supercomputing pilot.



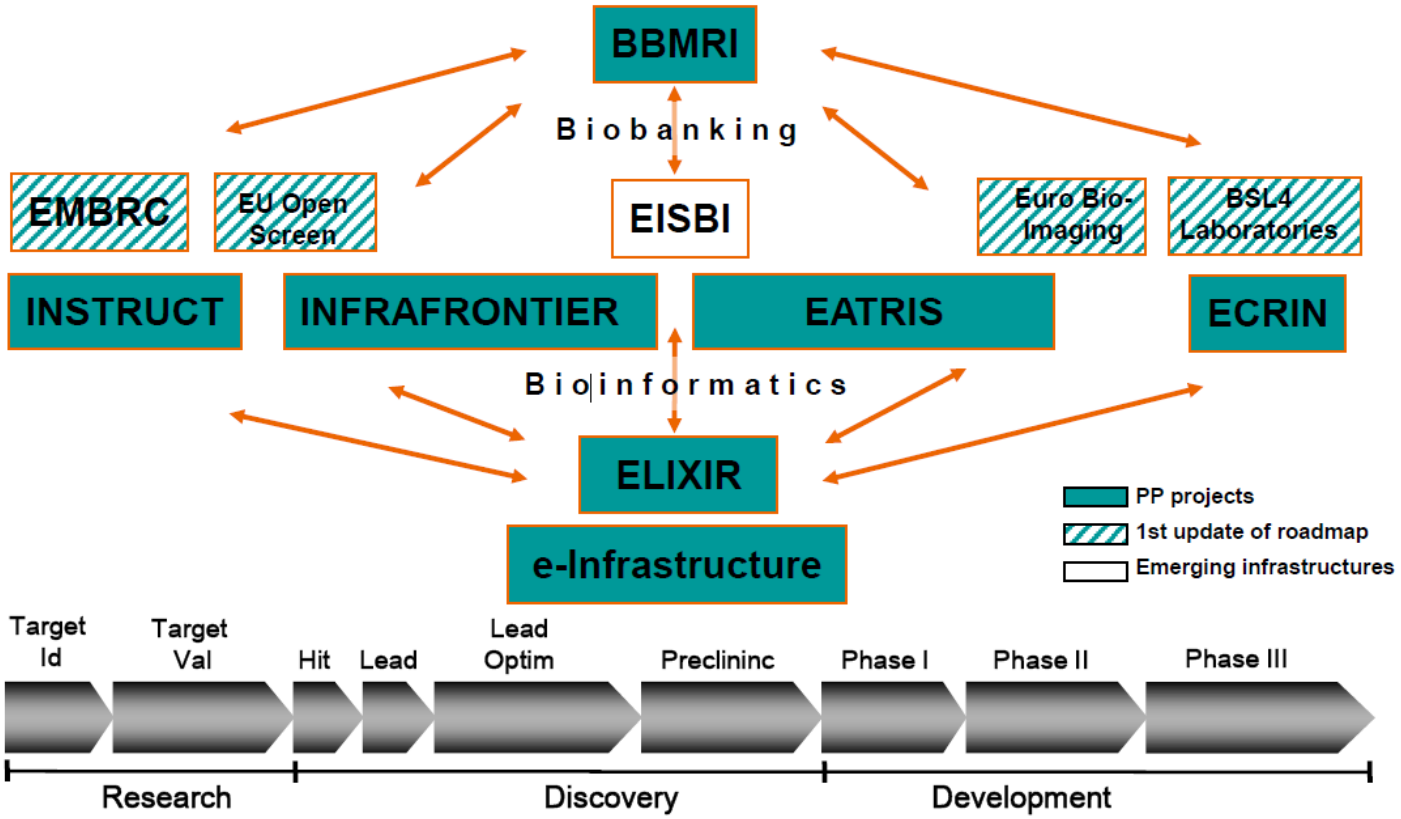
- Modus operandi of supercomputing centres does not naturally provide the 'services' needed for biology database production and serving.
- Processor farms dedicated to biology may be better solution
- Needs of biology are storage; large databases; simple repetitive searching over large databases
- Problems are probably more sociological than scientific.
- Way forward is not certain – ELIXIR Hub will use commercial Tier 3 data centre.

E-Infrastructure challenges.



- Current data are petabytes moving to exabytes over ESFRI period
- Data size comparable with largest research data generation projects
  - Data growth has been exponential for many years
  - Has exceeded improvements in CPU/disk/network
  - Data growth is accelerating
  - Tape backup no longer viable
  - Secure remote data replication required
- Biology has by far the largest research community
  - Maybe three million in Europe
- Community is used to ‘grid-on-demand’ free at point of delivery

Synergies between BMS RI



ELXIR Biology RI Interactions

Typically, an ESFRI BMS RI will interact with ELIXIR in at least one of the following ways:-

1. utilise one or more high-throughput technologies to **generate *large datasets*** that ELIXIR will be expected to accept, process, annotate (maybe with manual intervention), curate, archive and ***make available*** to the community
2. utilise one or more high-throughput technologies to **generate *large datasets*** that are out of scope for ELIXIR but whose full value can only be realised by ***linking*** them to data collections that are within the scope of ELIXIR.
3. generate ***large numbers of observations*** of biological entities that they will need to ***query*** against one or more large data collections that ELIXIR will be responsible for.
4. generate ***large numbers of observations*** of biological entities that they will need to ***query*** against one or more large data collections that are out of scope for ELIXIR but which will require ***standards and/or ontologies***, the development, maintenance and implementation of which will be in scope for ELIXIR.
5. other…

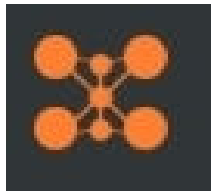
Preparatory phase activities

For each infrastructure we will need to

- Define the scope of the interaction between ELIXIR and the RI in question.
- Identify and catalogue the biological entities that will comprise the interaction.
- For each entity, identify and catalogue the relevant components that will be necessary to implement the interaction for that entity - components will include such things as data collections, standards and ontologies.
- Identify links, commonalities and other pragmatic considerations between the entities that will be necessary to enable an efficient and effective implementation.
- Create a generic requirement for the interaction based on the above.

NB. Each RI will have its own specific issues and requirements and the above will need to be tailored appropriately.

Construction phase activities



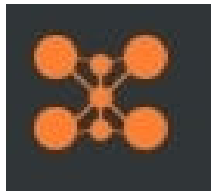
For each RI, based on work performed during Preparatory Phase, ELIXIR will

- Use Rapid Prototyping to develop a “quick-and-dirty” implementation to demonstrate feasibility.
- Identify potential users and ask them to evaluate the prototype.
- Use the prototype and the output of the evaluation to develop specific requirements.
- Identify the resources necessary to implement the specific requirements – this should include the means to implement any missing components and to modify or enhance any existing components as required.
- Where appropriate, and in association with the management of the RI in question, recommend the organisation or institution that would be best placed to implement the interaction.

Recent new activities at EBI for ESFRI RIs

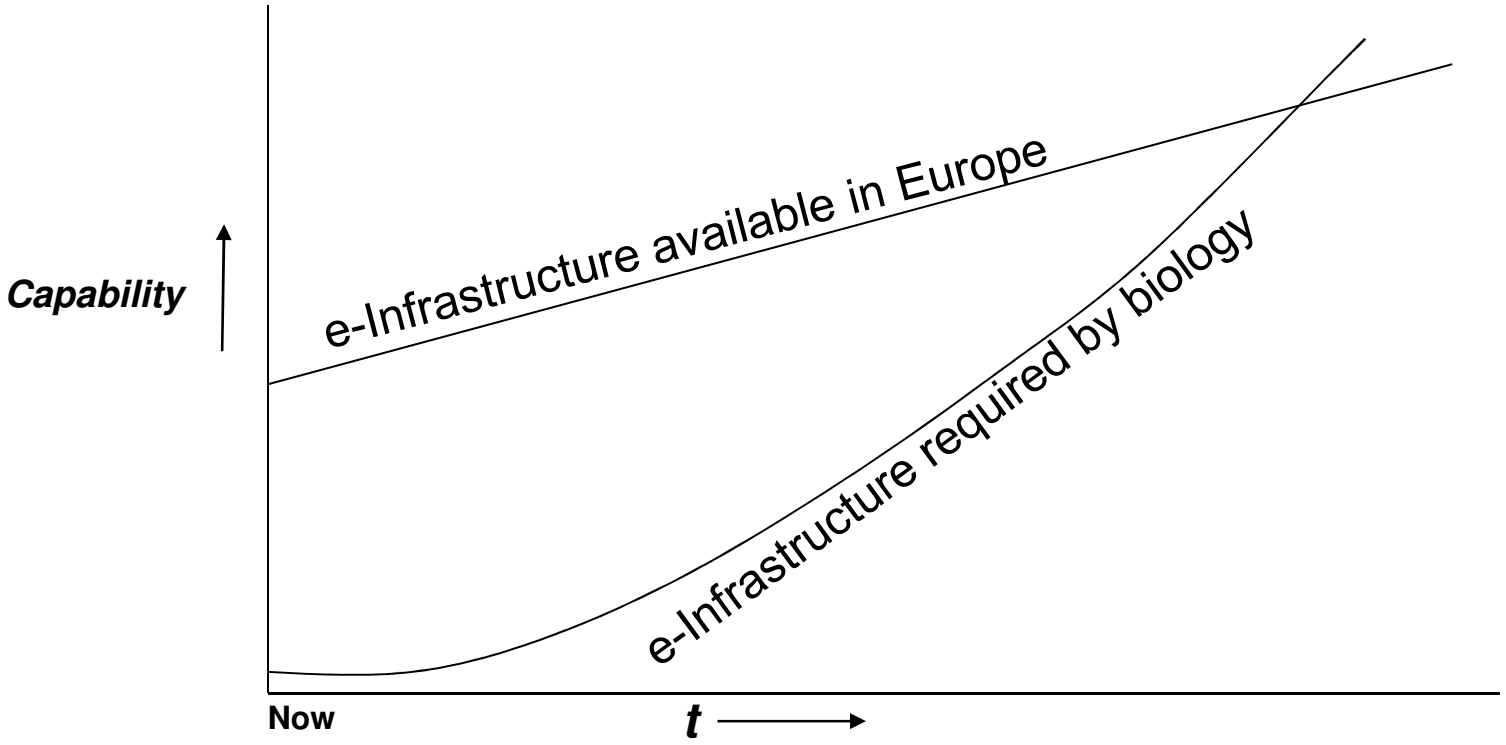
- Recovery of biomolecular data based on a specific sample identifier
  - BBMRI, EATRIS, ECRIN
- Recovery of biomolecular data based on a geographical location
  - LIFEWATCH, EMBRC
- Organism specific genome browsers
  - Infrafontier
- Chemoinformatics
  - EUOpenScreen

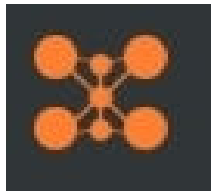
Summary



- Europe is facing unprecedented (grand) challenges.
- The solutions are (mainly) biological
- There are emerging (disruptive) technologies that offer ways forward
- These are very demanding of IT (e-Infrastructure)
- Biology has not been here before
- It will be necessary to move quickly to arrange the necessary funding streams
- Action is needed at every level (scientific, technical, funding, political, ...).

Biology e-Infrastructure Requirements





Thank you for your attention...