diXa and EUDAT

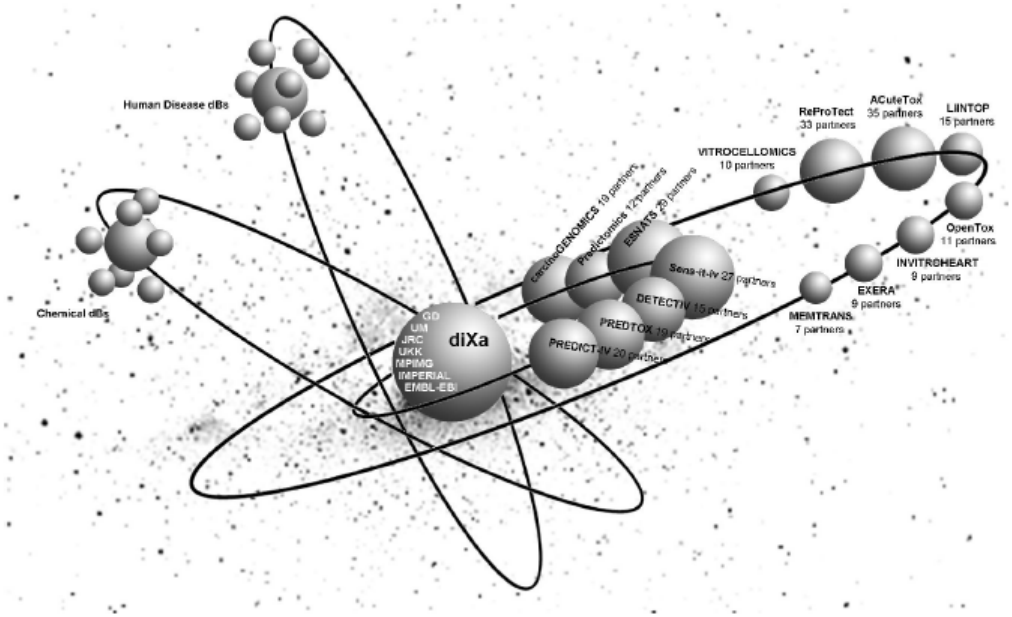
Safe replication service

- Ekaterina Pilicheva
- Ugis Sarkans
- European Bioinformatics Institute

EMBL-EBI

diXa data infrastructure for chemical safety

- **Grand vision** – replace animal-based chemical safety test models by non-animal assays (*in vitro* and *in silico*)
- **In practice** – web-based, open-access, and sustainable e-infrastructure for storing and searching data sets produced by past, current, and future EC research projects that target non-animal chemical safety tests



diXa partners

Partners

- **Maastricht University**
- **EMBL-EBI**
- **Genedata**
- **Max Planck Institute for Molecular Genetics**
- **Imperial College London**
- **Joint Research Centres (JRC)**
- **Klinikum der Universitaet zu Koeln**

EMBL-EBI

EMBL-EBI's mission

- Provide **freely available data** and bioinformatics services to all facets of the scientific community in ways that promote scientific progress
- Contribute to the advancement of biology through basic investigator-driven **research** in bioinformatics
- Provide advanced bioinformatics **training** to scientists at all levels, from PhD students to independent investigators
- Help disseminate cutting-edge technologies to **industry**
- Coordinate biological data provision throughout **Europe**

Data resources at EMBL-EBI

Genomes & variation

- Ensembl
- Ensembl Genomes
- Genome-phenome archive
- Metagenomics

Nucleotide sequences

- European Nucleotide Archive (ENA)

Literature & ontology

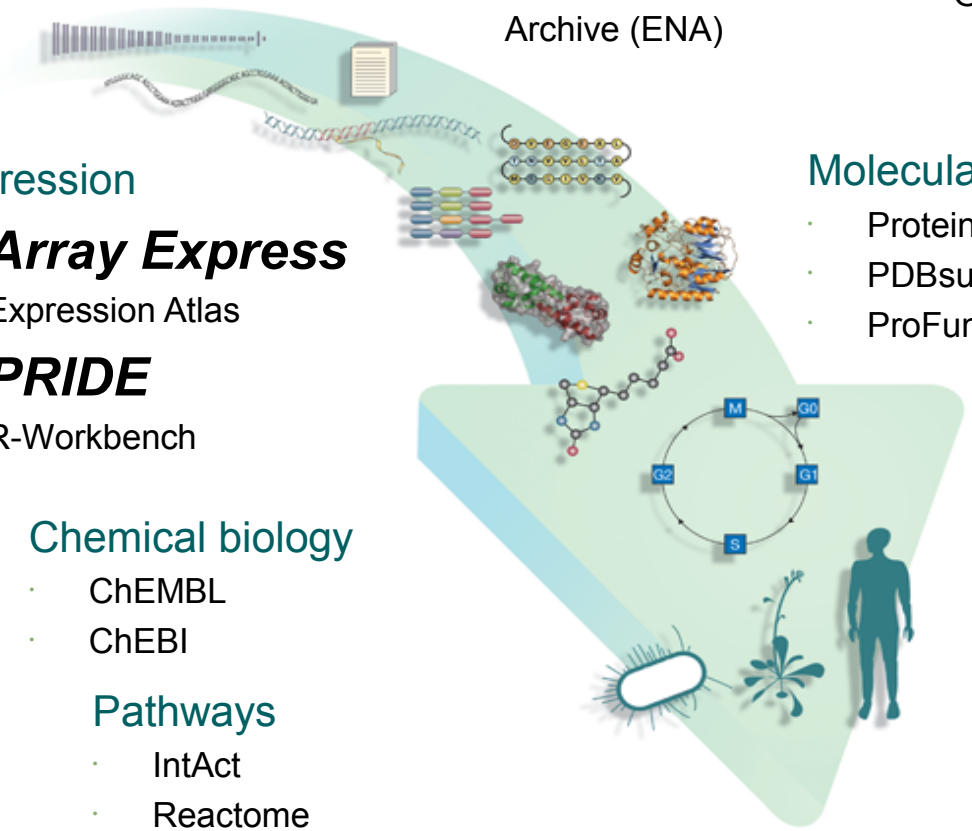
- Europe PubMed Central
- Gene Ontology

Expression

- ***Array Express***
- Expression Atlas
- ***PRIDE***
- R-Workbench

Molecular structures

- Protein Data Bank in Europe
- PDBsum
- ProFunc



Proteins

- The Universal Protein Resource (UniProt)
- InterPro

Chemical biology

- ChEMBL
- ChEBI

Patent sequences

- Non-redundant patent sequence dbs
- Patent compounds

Pathways

- IntAct
- Reactome
- ***Metabolights***

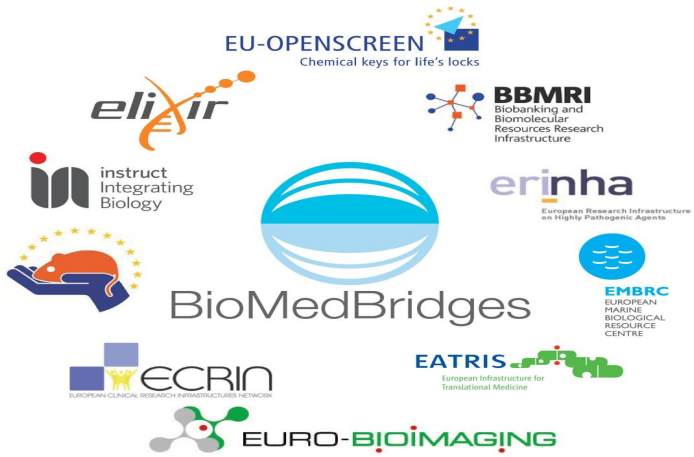
Systems

- BioModels
- Enzyme Portal
- BioSamples

Building data bridges from biology to medicine

BioMedBridges

- 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



diXa - EUDAT
Data Registration

motivation of the project for having the research data registered

- long-term archiving and data preservation
- easier communication with other domains
- to bring data closer to powerful computers for compute-intensive analysis

diXa - EUDAT
How diXa data organised?

is the data originally stored in data bases or in files? Can the data be distributed as sets of files?

Combination of files and databases (original data submissions – files, XMLdb stores metadata for GUI and search); usually there is a canonical file representation

what is the expected volume and granularity of data in the next 1-2 years ?

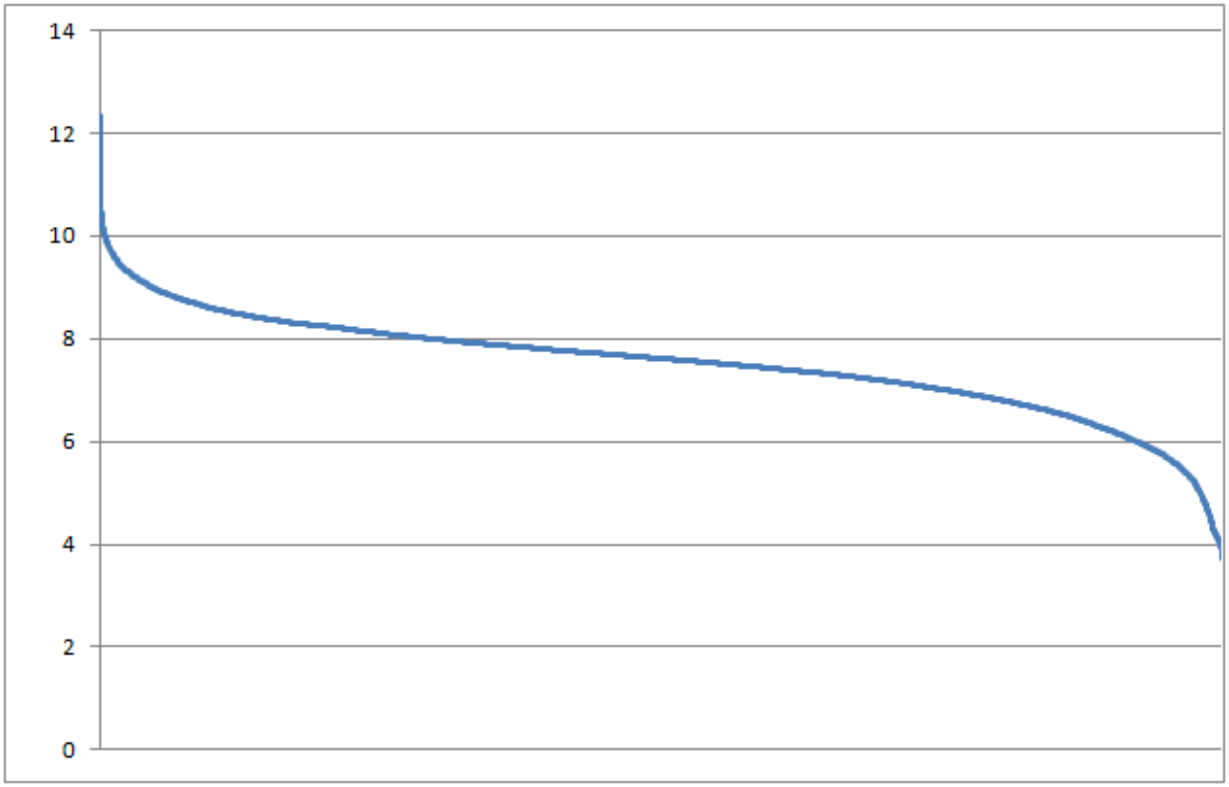
Up to 16TB in total range 10MB -1 TB

should the data be ingested at an EUDAT centre or should it be registered at the original site?

Initially at the EMBI-EBI

Data sizes illustration - ArrayExpress

- ArrayExpress functional genomics data repository
- ~35,000 datasets, total ~16Tb
- Size distribution (log10):



diXa - EUDAT
Safe Replication

- ***motivation***
  - several data replicas made available by different sites in Europe; data close to the centres with large compute capacity and bioinformatics expertise
- ***technical considerations***
  - data life cycle
    - updates – occasional, potential frequency – daily
- ***security/privacy***
  - public data - initial emphasis for SR
  - pre-publication data
  - human subject data – currently not considered