ELIXIR

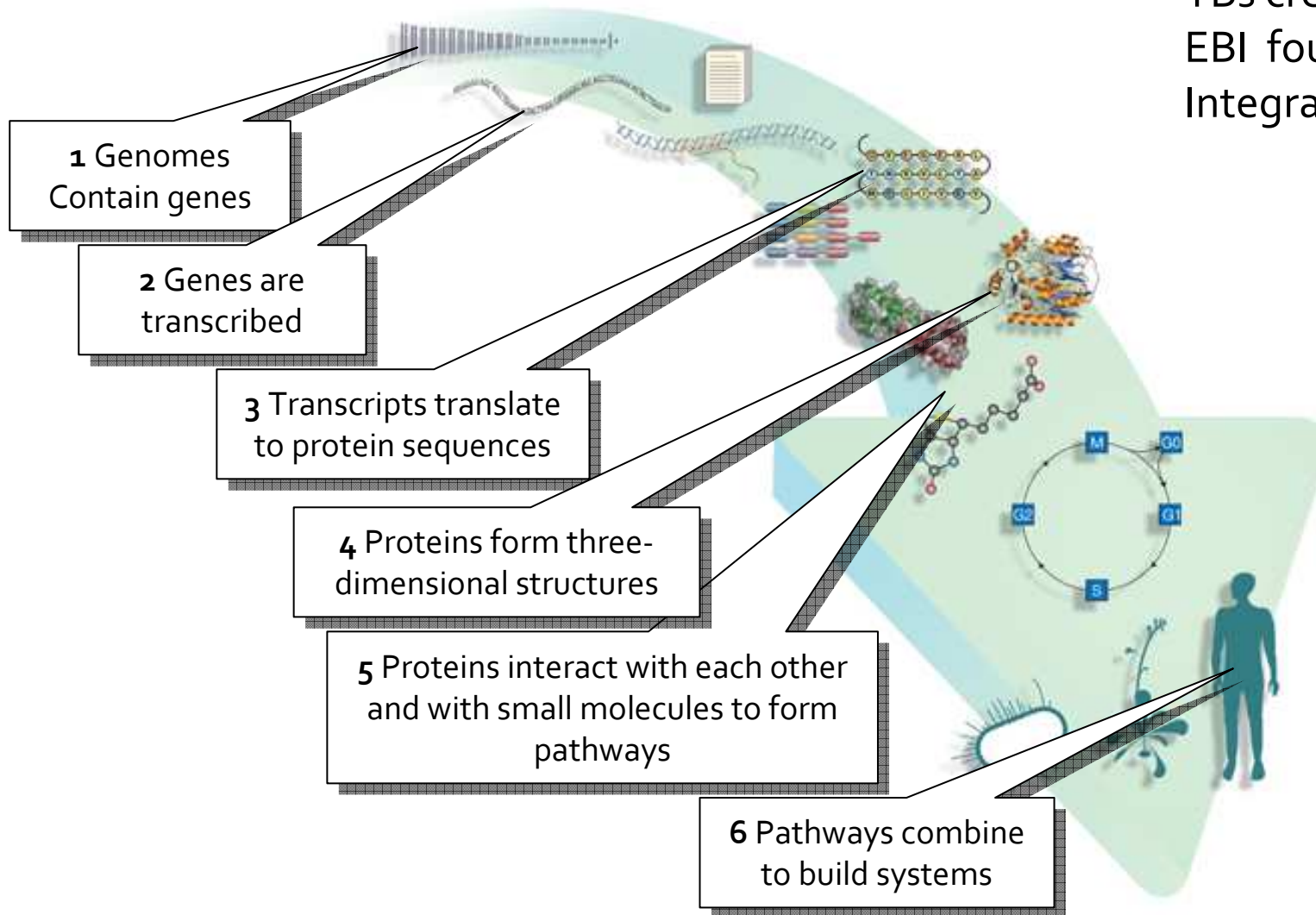## Europe's infrastructure for biological information

Life Science community presentation for EUDAT

European Life Sciences Infrastructure for Biological Information
www.elixir-europe.org

Bioinformatics underpins life-science

TBs created daily.
EBI founded 1993.
Integration!

1 Genomes
Contain genes

2 Genes are
transcribed

3 Transcripts translate
to protein sequences

4 Proteins form three-
dimensional structures

5 Proteins interact with each other
and with small molecules to form
pathways

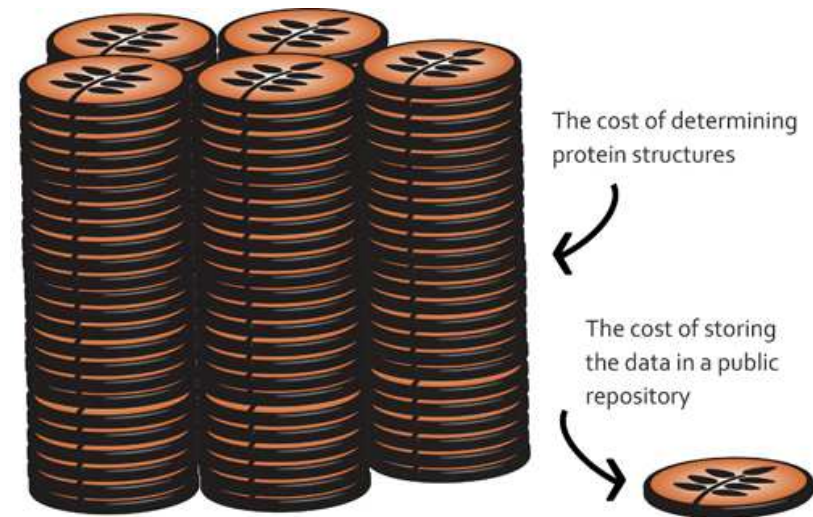**6** Pathways combine
to build systems

elixir

Data collections

- Ensembl: Joint project with Sanger Institute - high-quality annotation of vertebrate genomes
- Ensembl Genomes: Environment for genome data from other taxons
- 1000 Genomes: Catalogue of human variation from major World populations
- **EGA (European Genotype Archive): genotype, phenotype and sequences from individual subjects and controls**
- ENA (European Nucleotide Archive): all DNA & RNA, nextgen reads and traces
- ArrayExpress: Archive of transcriptomics and other functional genomics data
- Expression Atlas: Differentially expressed genes in tissues, cells, disease states & treatments
- UniProt: Archive of protein sequences and functional annotation
- InterPro: Integrated resource for protein families, motifs and domains
- PRIDE: Public data repository for proteomics data
- **PDBe: Protein and other macromolecular structure and function**
- ChEBI: Chemical entities of biological interest
- ChEMBL: Bioactive compounds, drugs and drug-like molecules, properties and activities
- IntAct: Public repository for molecular interaction data
- Reactome: Biochemical pathways and reactions in human biology
- Biomodels: Mathematical models of cellular processes
- GO: Gene Ontology, consistent descriptions of gene products
- CiteXplor: Bibliographic query system

Example: Protein Data Bank

- **Founded in 1971**

- Open access 3-D structural data of large biological molecules, such as proteins and nucleic acids

- Standard file format has been around for 40 years and the data collection, curation, annotation, integration and distribution is well organized and a daily process

- **International coordinated organization**. The four members (EU, USA, Japan) of wwPDB act as deposition, data processing and distribution centers for PDB data

The cost of determining protein structures

The cost of storing the data in a public repository

File formats, metadata descriptions, policies&agreements
http://www.wwpdb.org/docs.html

elixir

Example: European Genome-Phenome Archive

- All types of personally identifiable genetic and phenotypic data
  - E.g. Sequence reads from human cohort study to understand diabetes

- Data needs access control
  - **Who has access and to what purpose**

- One of the fastest growing EMBL-EBI resources
  - Studies
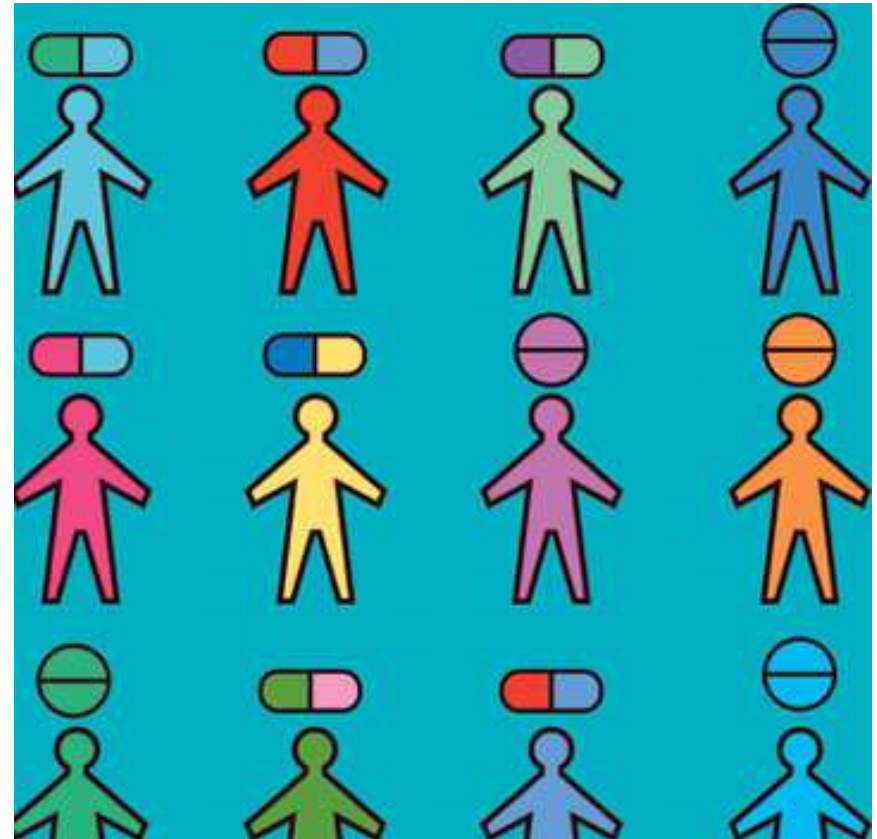 - Datasets
  - Data access committees
  - Data providers
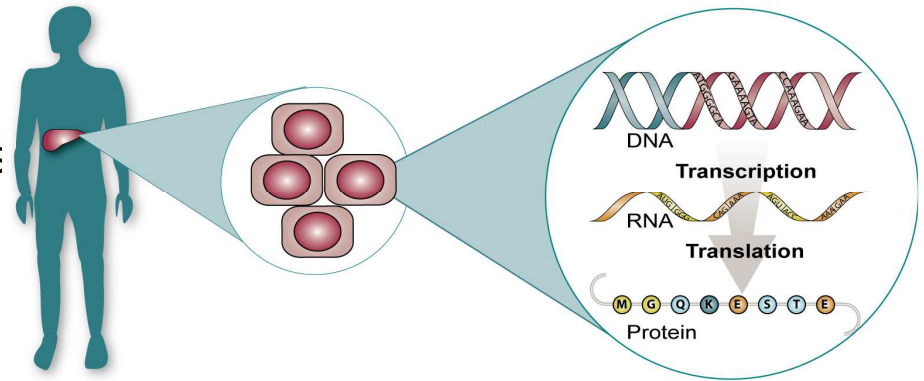
Process for data and metadata submission

https://www.ebi.ac.uk/ega/submission/manual

Personal genomes data

- Sequence data will not cross national boundaries.

- Every national health system will need **expertise to interpret it** and treat patients accordingly.

- Individuals need to be sure that their personal biological data are in safe hands.

The challenge

- Computer speed and storage capacity is **doubling every 18 months** and this rate is steady

- DNA sequence data is **doubling every 5 months** and this rate is increasing
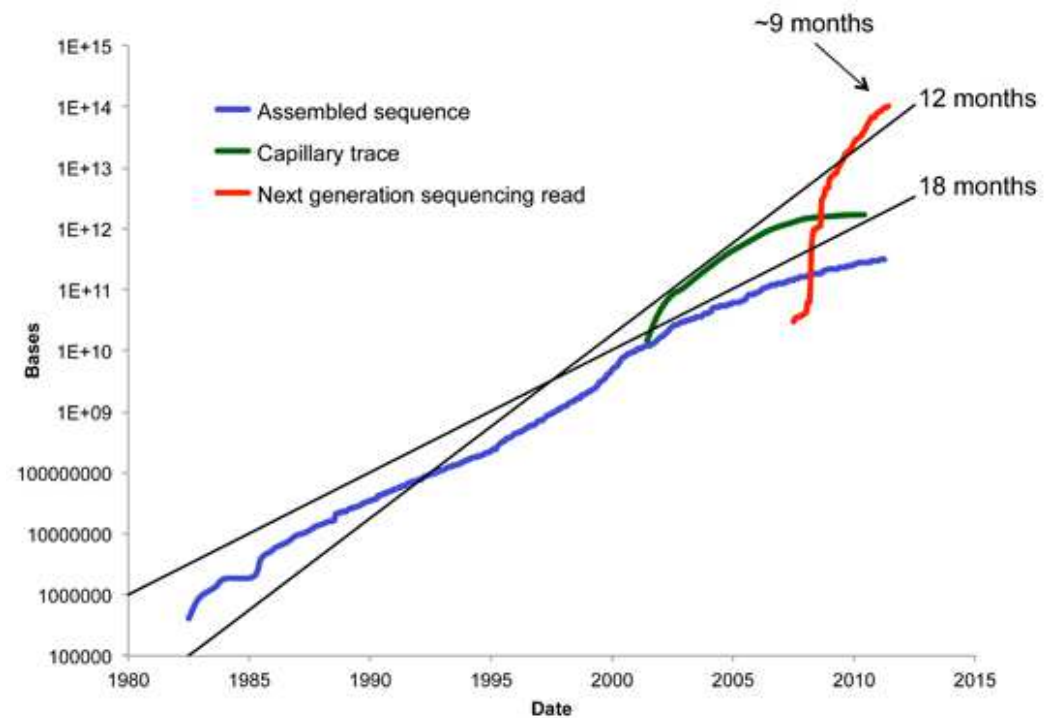

Guy Cochrane, ENA, EMBL-EBI

7

Biological data Service Provider

|	No access	Google, Twitter	Apple, Paypal (...rd)	Federated identities (& audit usage etc.)	... and 2nd access control (textmessage etc.)	Bank-style authentication	Very strong authentication
Open access data (hu...)				Maybe (depends on users)	No	No	No
Access to certain in... owner's w... (proprieta...)				OK	OK (overshoot?)	Maybe (costs SP?)	Maybe (overshoot?)
Access to data restricted due to contracts, regulations	No	No (who is responsible?)	Maybe	OK (access granted by SP)	OK (access granted by SP)	Maybe (cost?)	No (probably, not economical)
...	?	?	?	?	?	?	?
(Personal) medical data restricted by law	No	No	No	Maybe (access via ethical committees)	Maybe (access via ethical committees)	Maybe (probably no)	OK

Data service provider cannot support multiple ways

elixir
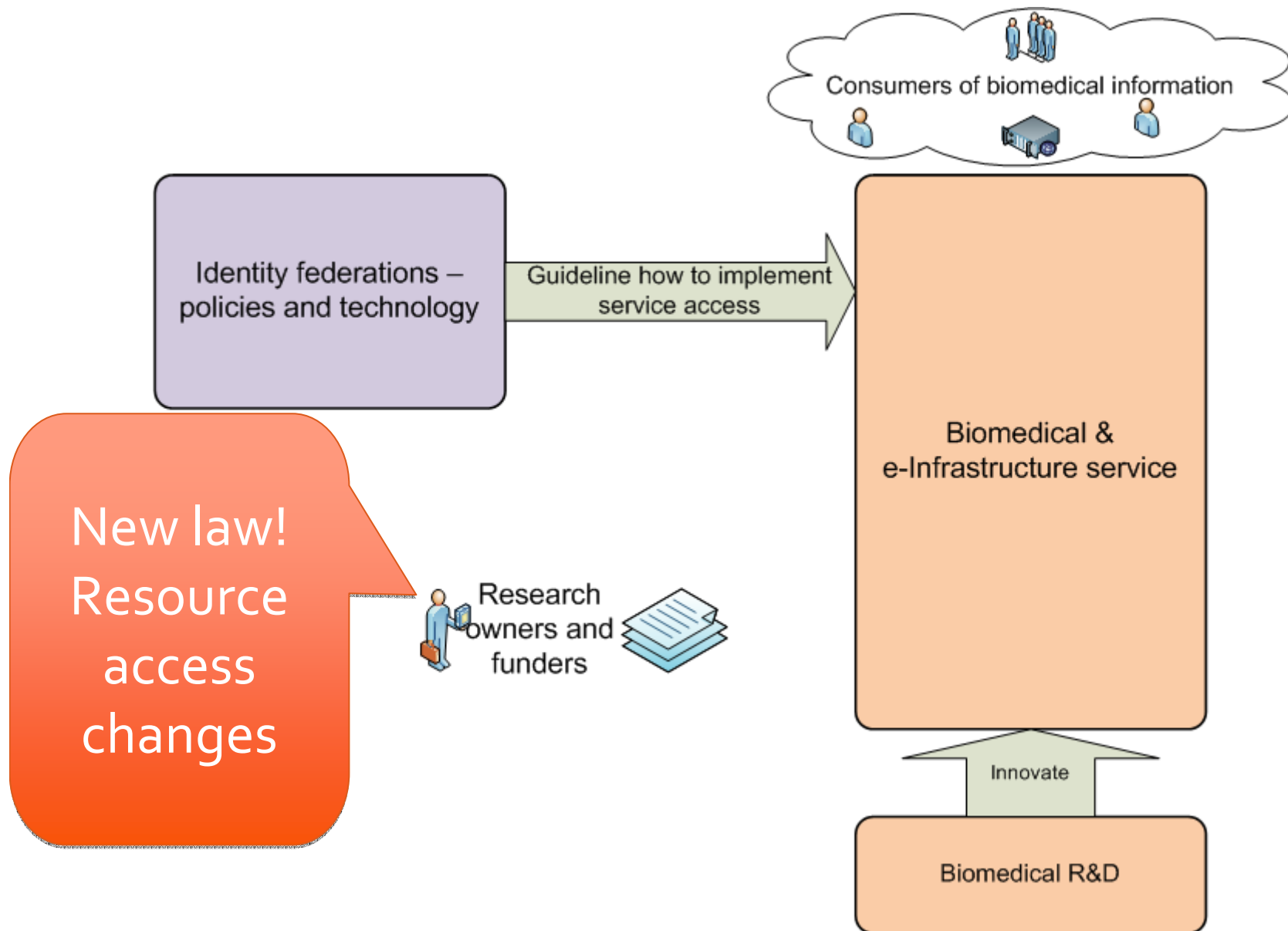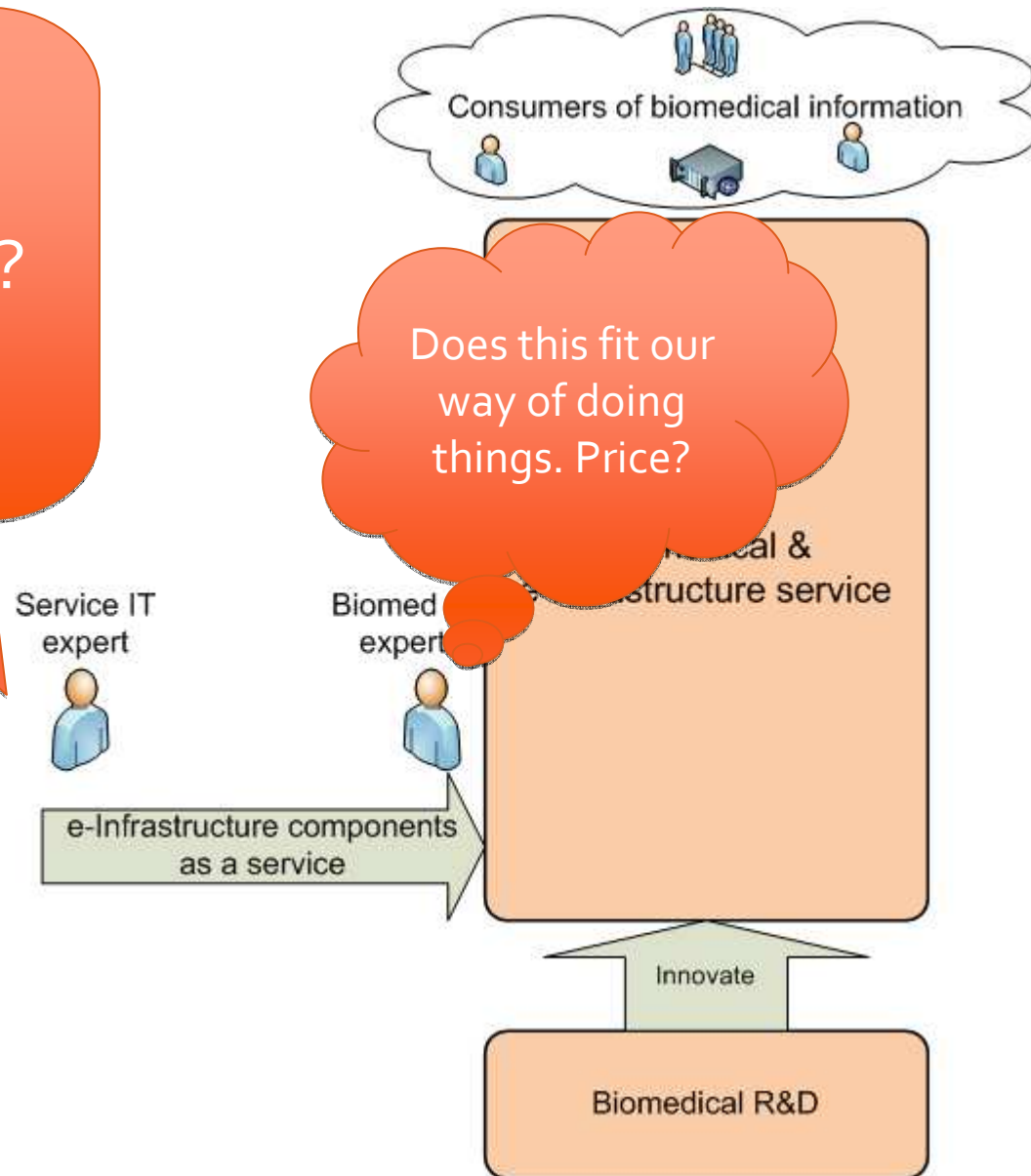
Teamplay of service&policy
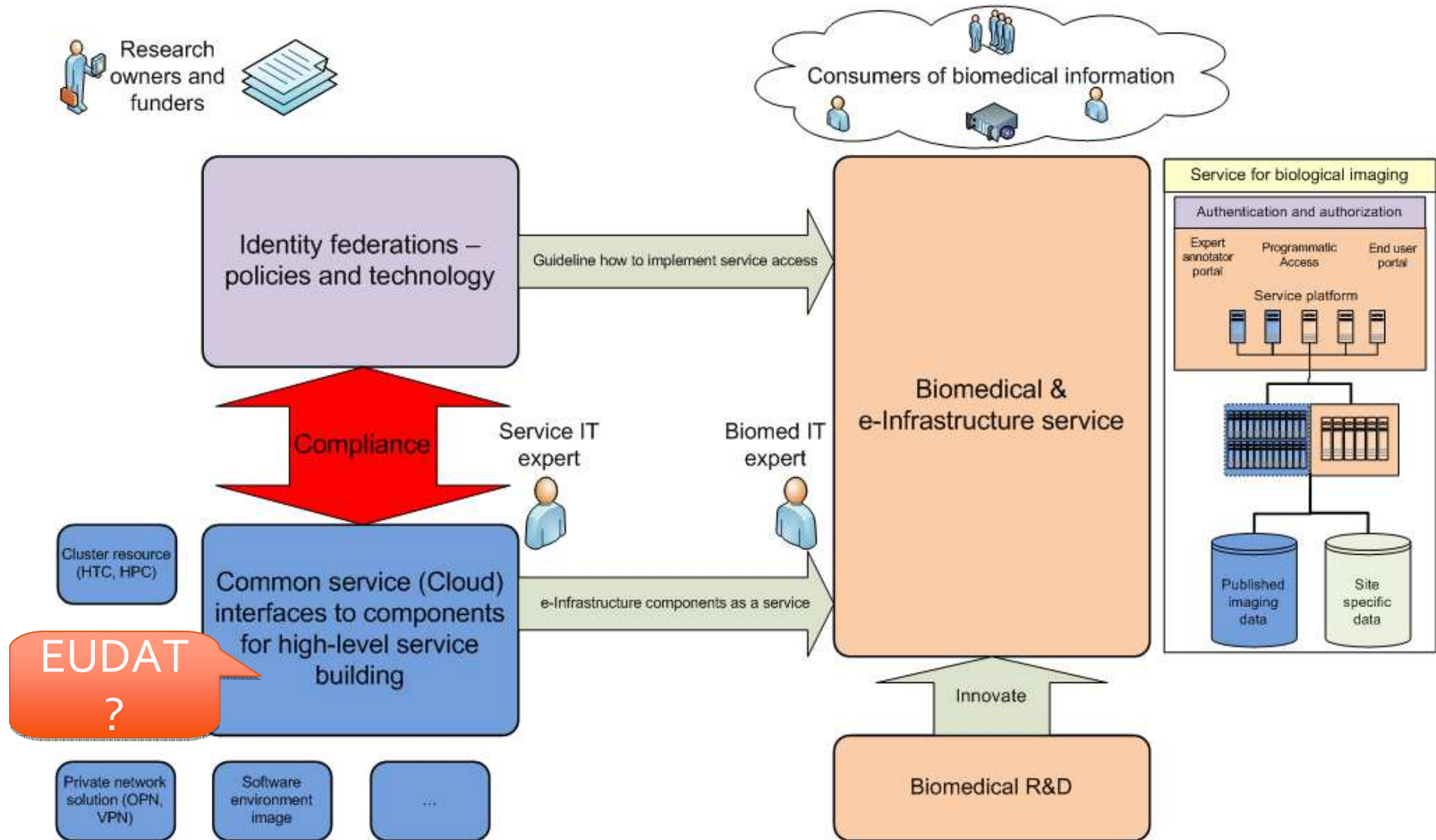
Suggestions to EUDAT

- EUDAT services storage, a recommended way is to deliver this so that it causes minimal change in the existing processes in Life Sciences
  - A good case to survey: imaging data creation and annotation
- For naïve users storage is where everything is taken care of for them e.g. YouTube
- For collaboration with RI's and experts like ELIXIR it is probably best to offer PaaS or IaaS and let that ESFRI's IT experts manage the community data to produce "consumer" services

elixir

Participate in BioMedBridges WP11: Technology Watch

- Contribute EUDAT expertise in the e-Infrastructure advisory of Biomedbridges EU FP7 project coordinated by ELIXIR
- Advisory comprises representatives of GÉANT, EGI.eu, PRACE & CERN + EUDAT as well as technical experts from the ESFRI BMS RIs
  - Bring together the technical experts of the BioMedBridges partners European ICT & e-Infrastructures to monitor and report on developments and provide advice to the project
  - Facilitate adoption of e-Infrastructure technologies by the BioMedBridges Work Packages and the ESFRI BMS RI
  - Communicate advice from the ICT Infrastructures and the e-Infrastructures to the BioMedBridges partners
  - Will produce annual reports on status of e-Infrastructures relevant to the progress of BioMedBridges including requirements for use cases & recommendations for adoption of new technologies and standards.

Summary from ELIXIR to EUDAT

The purpose of ELIXIR is provide

"a sustainable infrastructure for biological information in Europe."

- ELIXIR is motivated by EUDAT "data proliferation", but ELIXIR also needs
  - Mechanisms to share the task and the cost between all of the member states of the EU: Hub and Nodes
  - Collect and Integrate many diverse data-types medicine, agriculture, biotechnology, etc : this requires deep understanding of biology
  - Deploy the data collections to a very large, diverse and fast-growing community of users: at least 3 million consumers now
  - Tailor the data collections to the needs of European Industry, in particular Pharmaceuticals and Biotech
  - Support the data collections and services with comprehensive training programmes

Informed: ELIXIR eNewsletter

- New quarterly eNewsletter for ELIXIR

- Will provide the latest news and updates in relation to the ELIXIR project

Sign up for *Informed* at the following:

http://eepurl.com/iquj-/