ELIXIR

Rafael C Jimenez

ELIXIR CTO

- Introduction to ELIXIR
- Challenges in life sciences
- ELIXIR Technical activities
- Collaborations with e-infrastructures

European Life Sciences Infrastructure for Biological Information

www.elixir-europe.org

ELIXIR

Sustainable European infrastructure for **biological research data**

Facilitate research

Safeguard data and build **sustainable data services**

ELIXIR Members

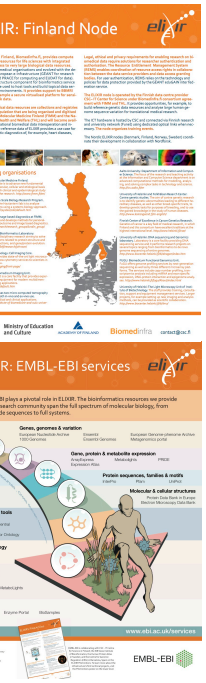
Connects **national centers** and **EMBL-EBI**

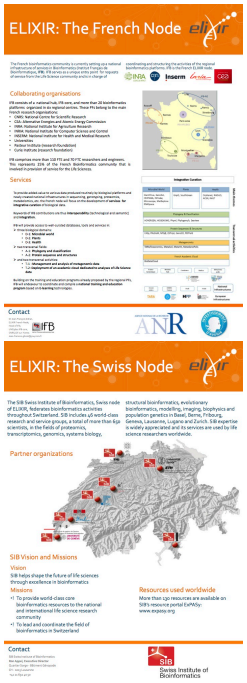
Participated by major bioinformatics service providers (**~130**) and supported by **17 EU member states**

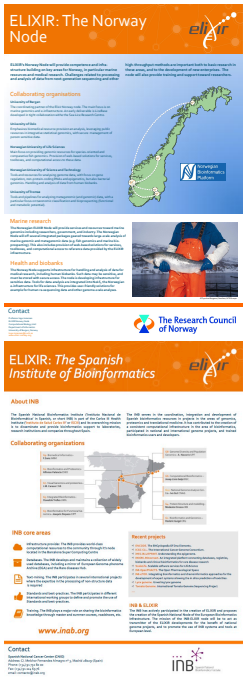
- 11 Members
  - *Czech Republic, Estonia, Denmark, Finland, Israel, Netherlands, Norway, Portugal, Switzerland, Sweden, UK*
- 6 Observers
  - *Belgium, Greece, France, Italy, Slovenia, Spain*

ELIXIR node proposals

ELIXIR deliver services through **ELIXIR Nodes** building on national strengths and priorities







ELIXIR Hub drives coordination

ELIXIR strategic drivers

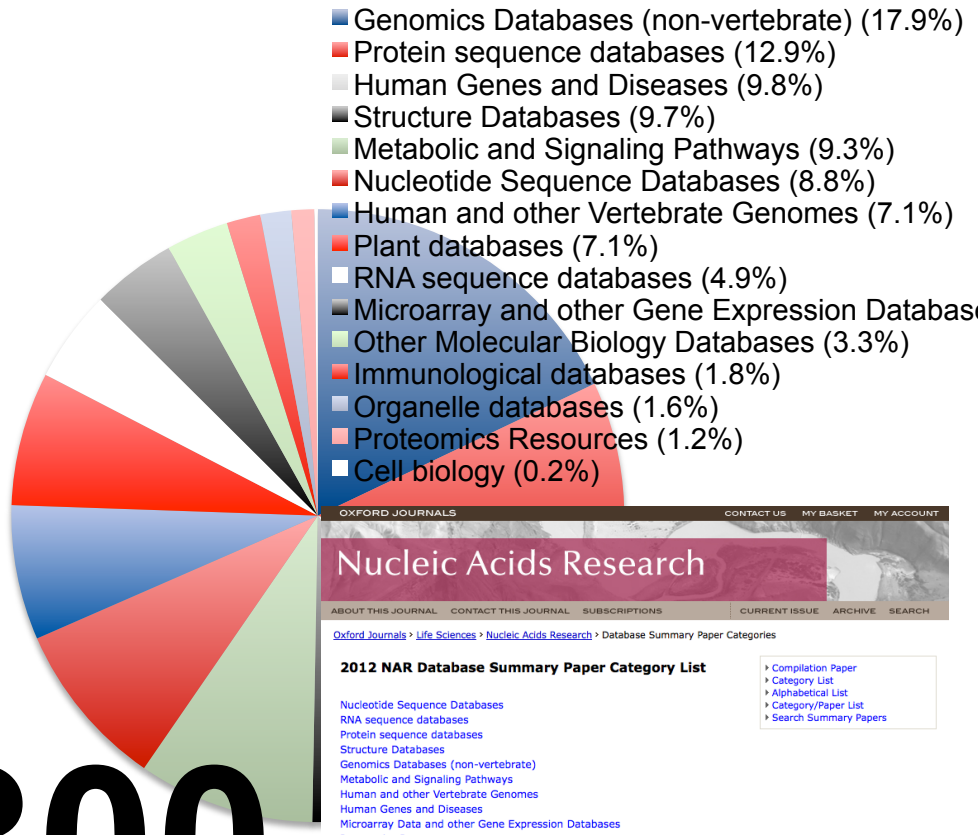
1. Scale with the challenge of **data growth**
2. Secure and deliver the **core data resources**
3. Provide **discoverable tools**, services and connectors to drive data access and exploitation
4. Provide robust **technical platforms** for secure data access, data exchange and compute
5. Develop and maintain **standards** for data management, reuse and integration
6. Drive partnerships with **user communities** and other organisations to ensure high impact
7. Close the computational biology skills gap through a comprehensive **training** programme for professionals
8. Support **innovation** in big data biology

Challenges in life sciences

www.elixir-europe.org

Data resources in life science

- Many
- Diverse
- Disperse



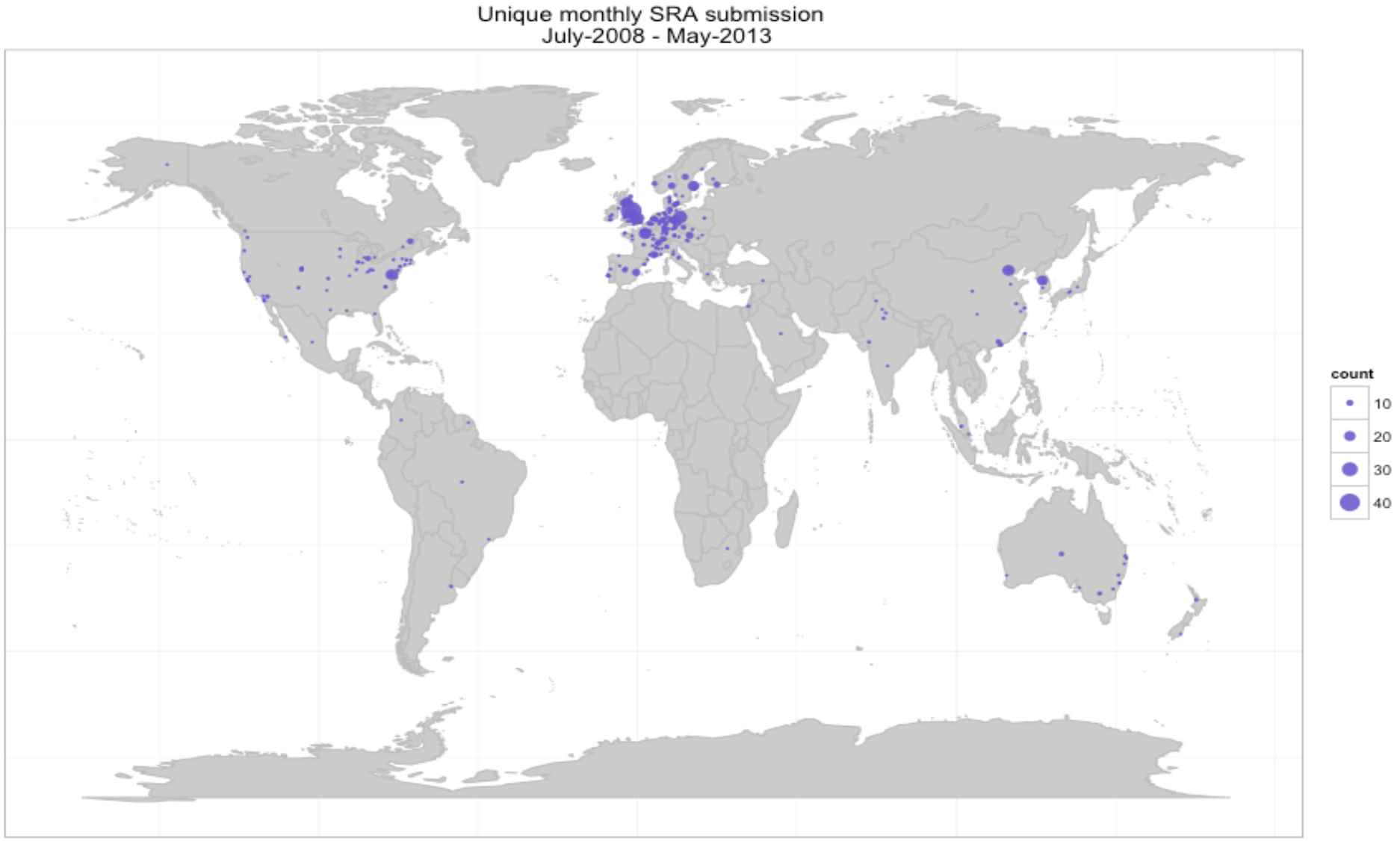
~1800
molecular biology data resources

Nucleic Acids Research annual Database Issue
and the NAR online Molecular Biology Database Collection in 2012.
MY Galperin, GR Cochrane – Nucleic Acids Research, 2011

elixir

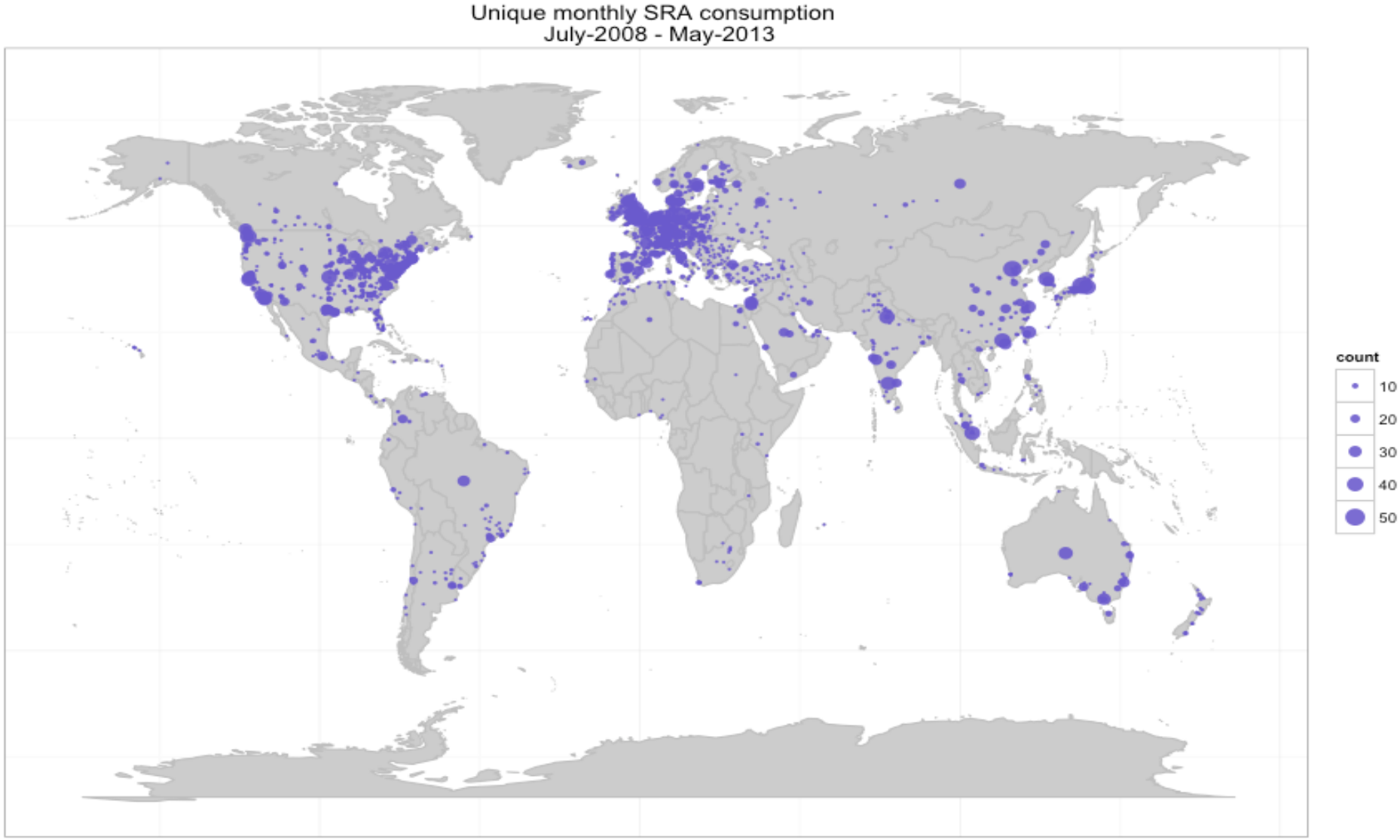
Disperse science

Data production

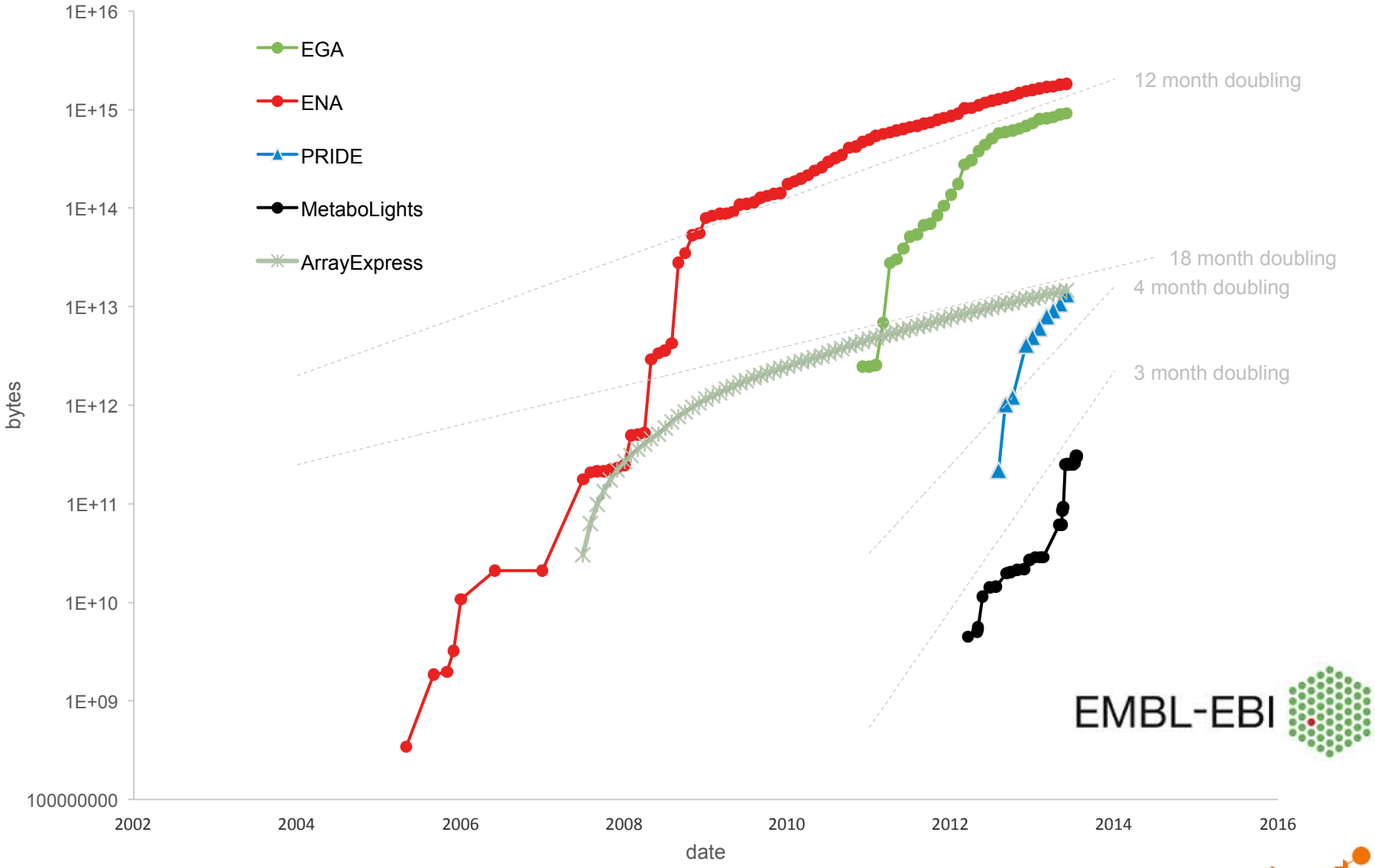


Disperse science

Data consumption



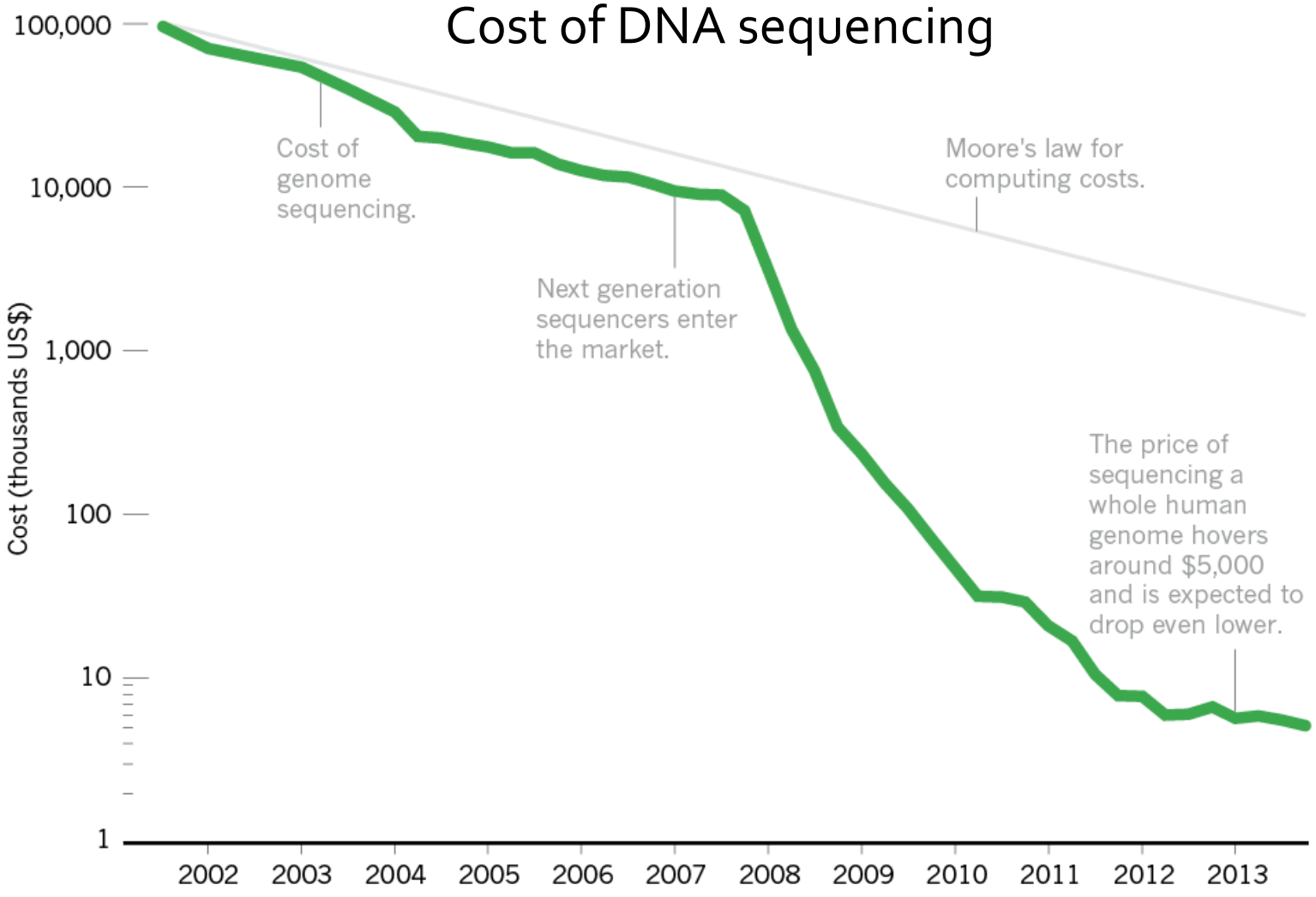
Data growth



EMBL-EBI

elixir

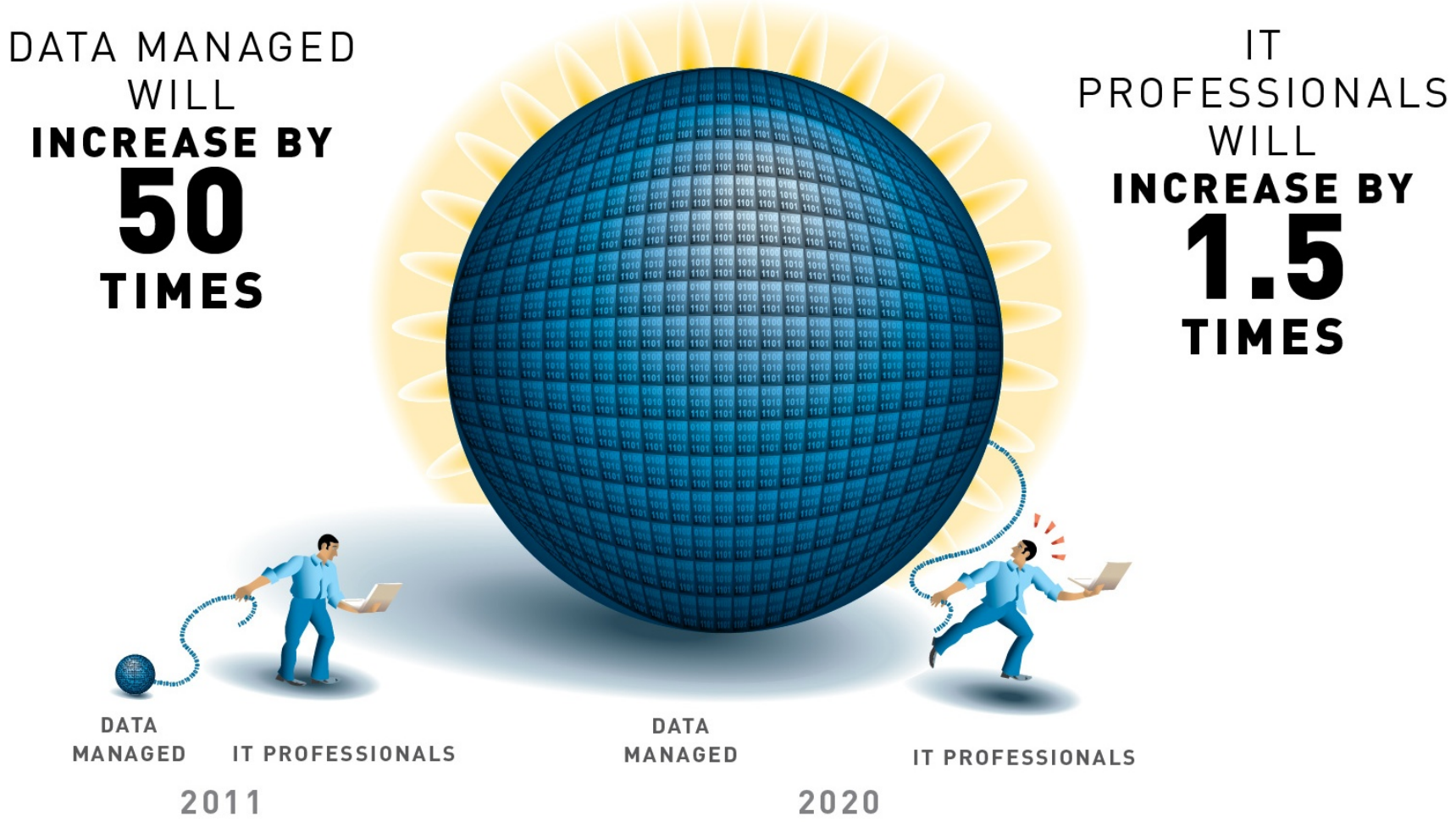
Cost of data production technologies declines faster than storage



Data generation vs. data transfer

		Network File Transfer		
24 hours		**1 Gb**	**100 Mb**	**10 Mb**
	DNA sequencing ~100 GB	~30 min	~5 hours	~2 days
	Mass spectrometry ~4 TB	~9 hour	~4 days	~5 weeks
	Microscopy ~4 TB	~9 hour	~4 days	~5 weeks

Shortage of professionals



THE 2011 IDC DIGITAL UNIVERSE STUDY
sponsored by EMC

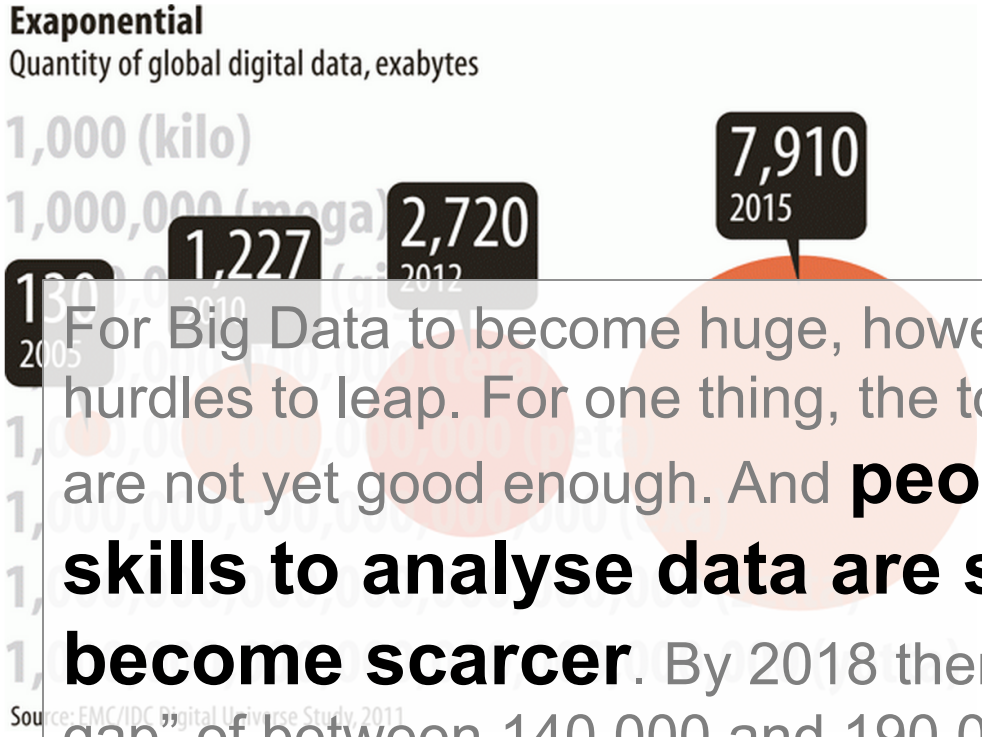
Business

Welcome to the yotta world

Comment (1) | Print | E-mail | Reprints & permissions

Big Data will flood the planet

Nov 17th 2011 | From The World In 2012 print edition



Source: EMC/IDC Digital Universe Study, 2011

Even if you still have to think twice about the meaning of "giga" and "tera" in computer-speak, you'd better get ready for "peta", "exa" and "zetta". These binary prefixes, which

For Big Data to become huge, however, there are still hurdles to leap. For one thing, the tools to analyse data are not yet good enough. And **people with the skills to analyse data are scarce and will become scarcer**. By 2018 there will be a "talent gap" of between 140,000 and 190,000 people, …

Follow *The Economist*

Latest blog posts - All times are GMT

The Economist explains: How America defines religious freedom

Poland's agriculture: A golden age for

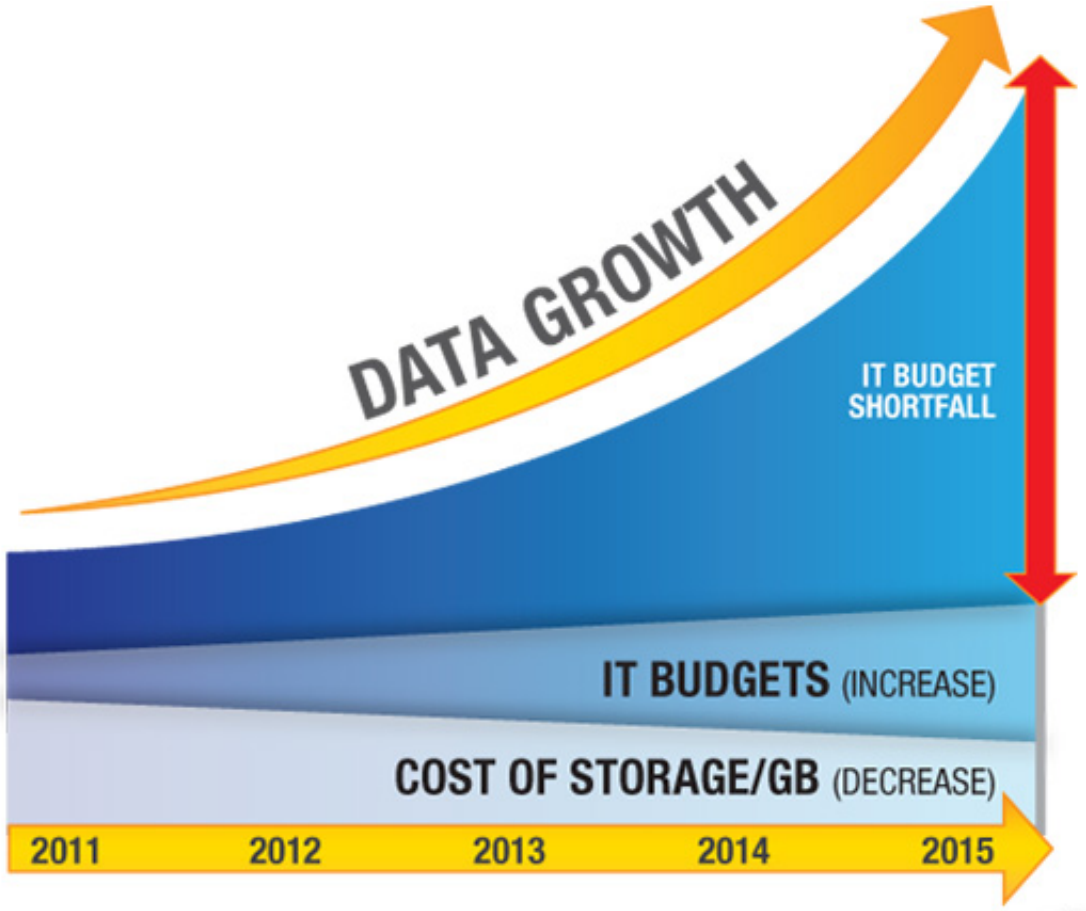
The enigma of flight 370: Dashed hopes
Newsbook - Mar 24th, 17:56

Feedback

Challenges

- **Sustain** data and services
- Make data and service **interoperable**
  - Necessary to integrate data
  - Specially medical, clinical and research
- Data too big to **store**, **exchange** & **compute**? Forthcoming challenges …
  - Data production grows faster than storage
  - Cost of data production technologies declines faster than storage
  - It takes longer to transfer data than produce the data.
- Privacy, security & access (AAI)
- Training

How to reduce the IT budget shortfall?



http://www.eweek.com/

BioMedBridges workshop

E-Infrastructure support for the life sciences:

Preparing for the data deluge

15 May, 2014

Genome Campus, Hinxton, UK

BioMedBridges training workshop

Data management for research infrastructure **project managers**

16-17 or 18-19 February 2015, Munich

to be confirmed

Stephanie Suhr

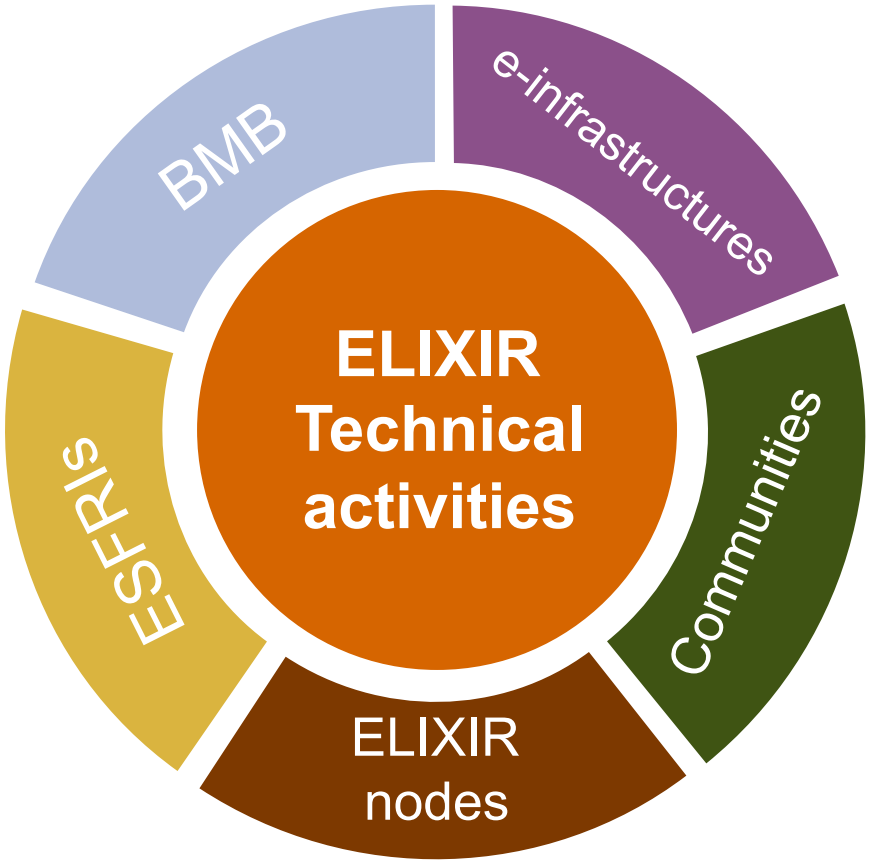
ssuhr@ebi.ac.uk

Technical activities

www.elixir-europe.org

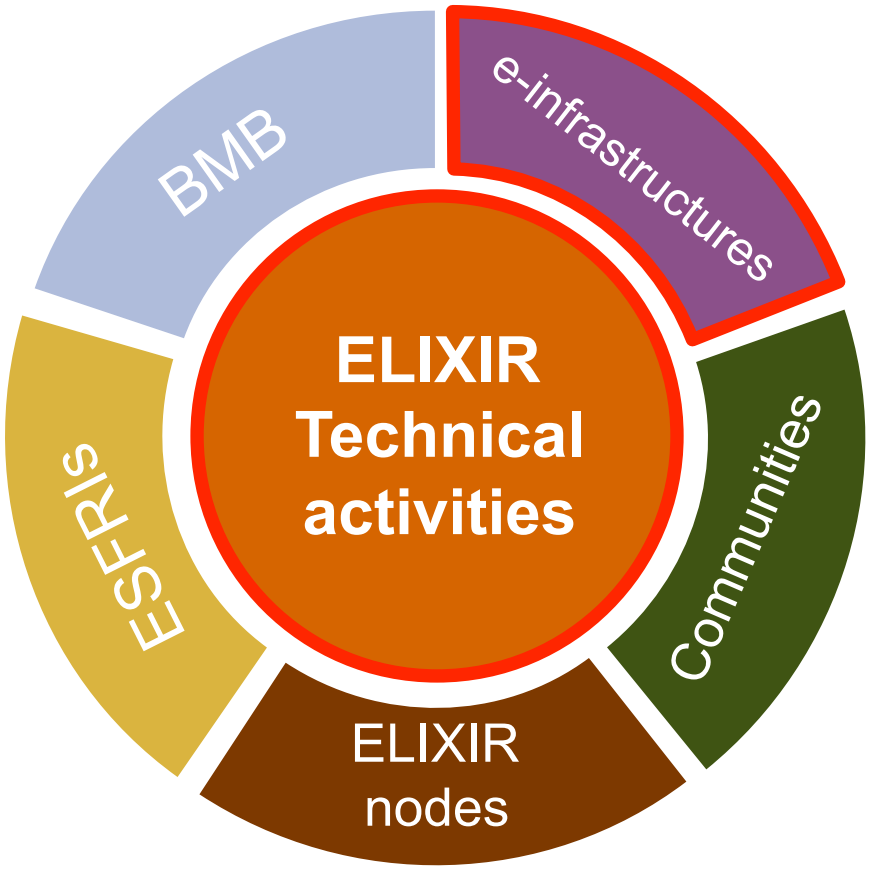
ELIXIR technical activities

- ELIXIR Node activities, Task forces, Pilots
- Technical activities among different interest groups …



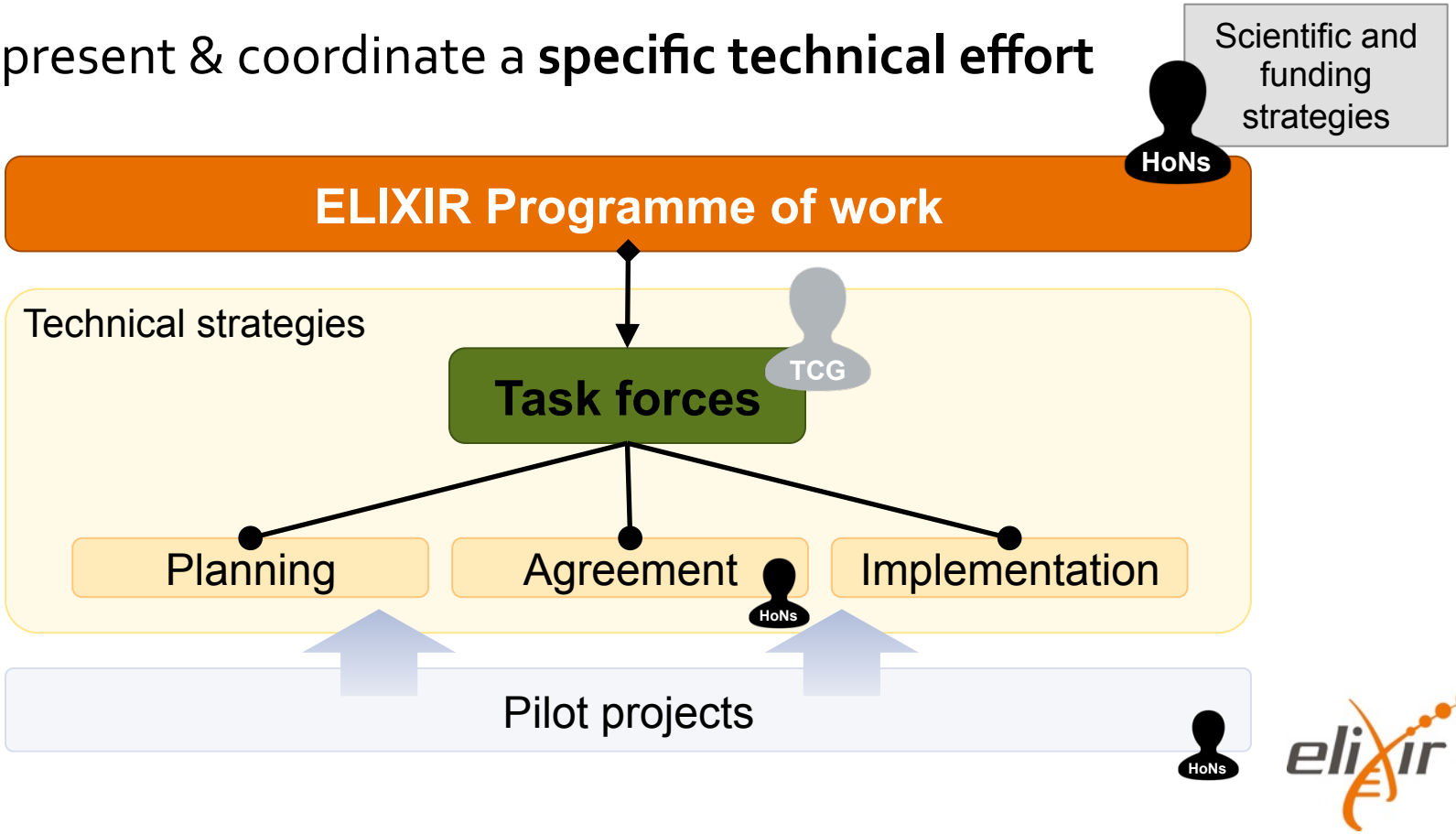
ELIXIR technical activities

- ELIXIR Node activities, Task forces, Pilots
- Technical activities among different interest groups …



Task forces

- **Short term** working groups
- **Driven** by ELIXIR **technical coordinators**
- **Participated by experts**
- Represent & coordinate a **specific technical effort**



ELIXIR Task Forces

Programme of work	Task forces
Technical services	Cloud
	Storage
	Authentication and authorization
Tools Interoperability and Service Registry	**Service registry**
Data resources and services	**Metrics, monitoring & quality control**
Management and Operations	**Communication**
	Website
Training	Training portal
	e-Learning
Data interoperability, vocabulary and ontology services	Interoperability

ELIXIR Task Forces

Programme of work	Task forces
Technical services	Cloud
	Storage
	Authentication and authorization
Tools Interoperability and Service Registry	**Service registry**
Data resources and services	**Metrics, monitoring & quality control**
Management and Operations	**Communication**
	Website
Training	Training portal
	e-Learning
Data interoperability, vocabulary and ontology services	Interoperability

Cloud TF

Steven Newhouse & Mirek Ruda
cloud-tf@elixir-europe.org

Use cases

- **Infrastructure as a Service (IaaS)**
- **Platform as a Service (PaaS)**
- **Virtual Machine Repository or Marketplace Portal**
- Virtual Clusters
- Running Data Analysis Pipelines
- Data Extraction
- Scalable Web Service Hosting
- Shared Environment
- Virtual Desktops for Immediate Use
- Software Development and Testing
- Appliance

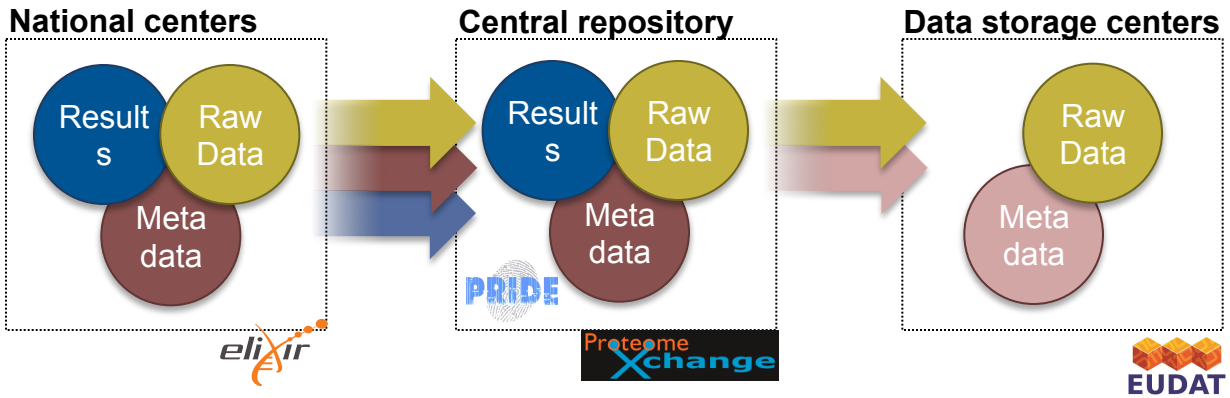
Projects

- EMBASSY cloud
- EGI - ELIXIR Competence Centre
- EGI Federated Cloud
- Cloud portal in collaboration with Bull
- Virtual Machine Library
- Australian Genomics Workbench
- Bioinformatics Tools: Genome Analysis and Protein Folding (GAPF)
- ELIXIR ENSEMBL replication Pilot
- Cross-site VM Operation - EYRG
- BBMRI Cloud
- eLearning pilot with Computing pilot

Storage TF

Mikael Borg & Irene Nooren
storage-tf@elixir-europe.org

- Survey of storage practices in node countries
- Discussions with EUDAT and EGI
- Integrating ELIXIR reference datasets within the European Grid Infrastructure
  - Facilitate replication & discovery of replicated data
- List of use cases
  - BioImaging (collaboration agreement)
  - Proteomics (pilot)



AAI TF

Mikael Linden & Ilkka Lappalainen
aai-tf@elixir-europe.org

Use case	Identity and authentication	Authorisation	Access control enforcement
1: Guest account service	x		
2: Bona fide researcher service		x	
3: Availability service			x
4: REMS		x	
5: EGA			x
6: Cloud and EGA			x
7: Service Provider			x

8. Attributes across services
9. Sharing generic web based resources
10. eLearning platform and Cloud

ELIXIR Pilot Projects

1. **ELIXIR Facing Cloud Support and Virtual Machines - with SIB**
2. **ELIXIR Data IO to pilot the continuous transfer of major archive resources to a remote European location - with CSC, Finland**
3. **Establishing EGA Distributed authentication - with CSC, Finland**
4. Establishing EGA as joint venture – with CRG, Spain
5. Improving links between Human Proteome Atlas (HPA) and EMBL-EBI resources
6. **BILS-ProteomeXchange integration using EUDAT resources**
7. Interoperable controlled-access big data transfer technology for ELIXIR - application to EGA EBI / CRG ELIXIR collaboration and beyond
8. Harmonising Marine Metagenomics pipelines

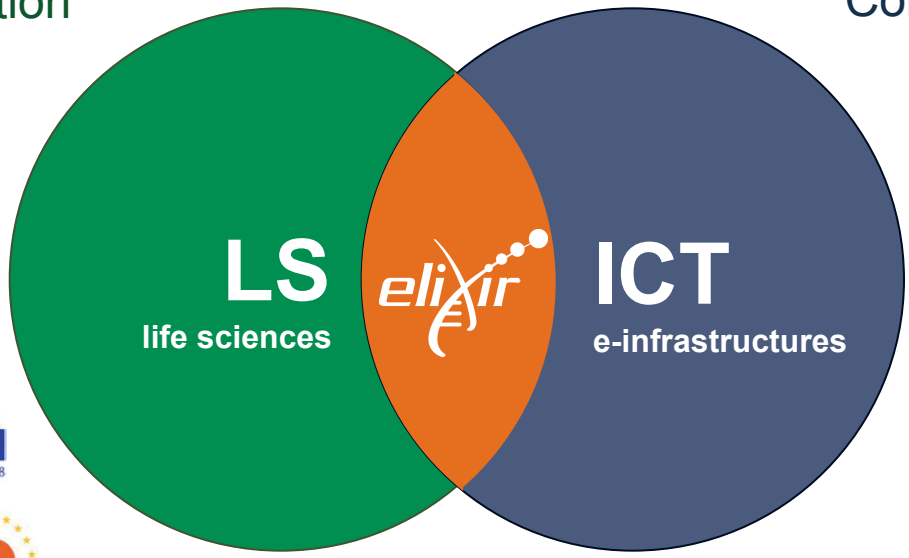
Collaborations with e-infrastructures

www.elixir-europe.org

Research infrastructures

Facilitate research

Scientific information
Physical facilities

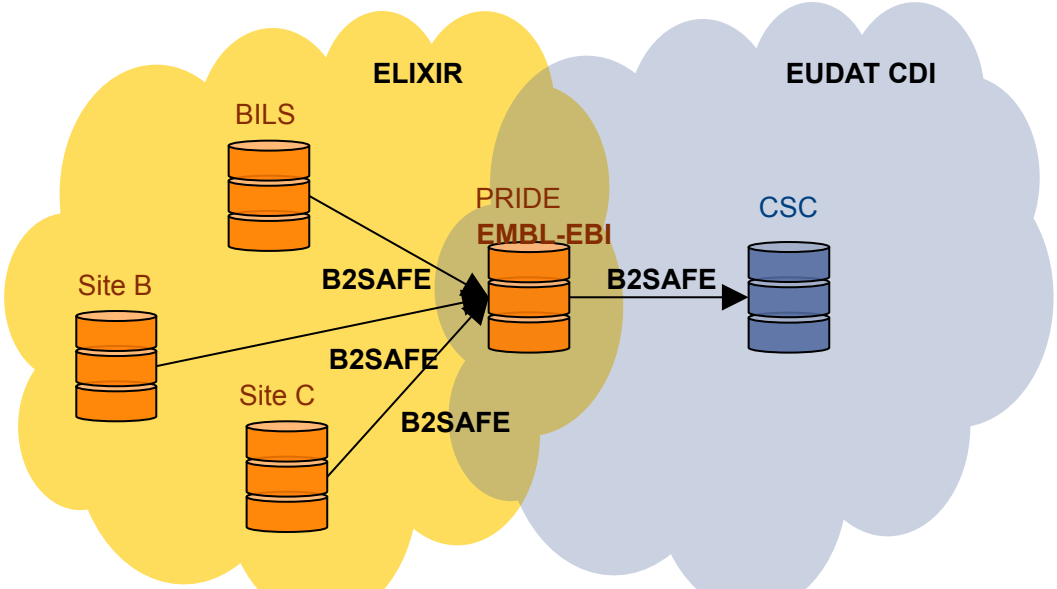


Computation
Transfer
Storage

EURO-BIOIMAGING
ECRIN
EMBRC
EUROPEAN MARINE BIOLOGICAL RESOURCE CENTRE
EU-OPENSCREEN
ESFRI ROADMAP 2008
EATRIS
European Infrastructure for Translational Medicine
instruct
Integrating Biology
BBMRI
Biobanking and Biomolecular Resources Research Infrastructure
PRACE
EGI
GÉANT
EUDAT

Collaborations with EUDAT

- EUDAT minimum metadata for life science domain (Hub)
  - Describe EUDAT data sets with minimum metadata
- EUDAT FAIR data (NL)
  - Cross Research Infastructure interoperability of research data
- Proteomics repositories integration using EUDAT resources (SE & EMBL-EBI)

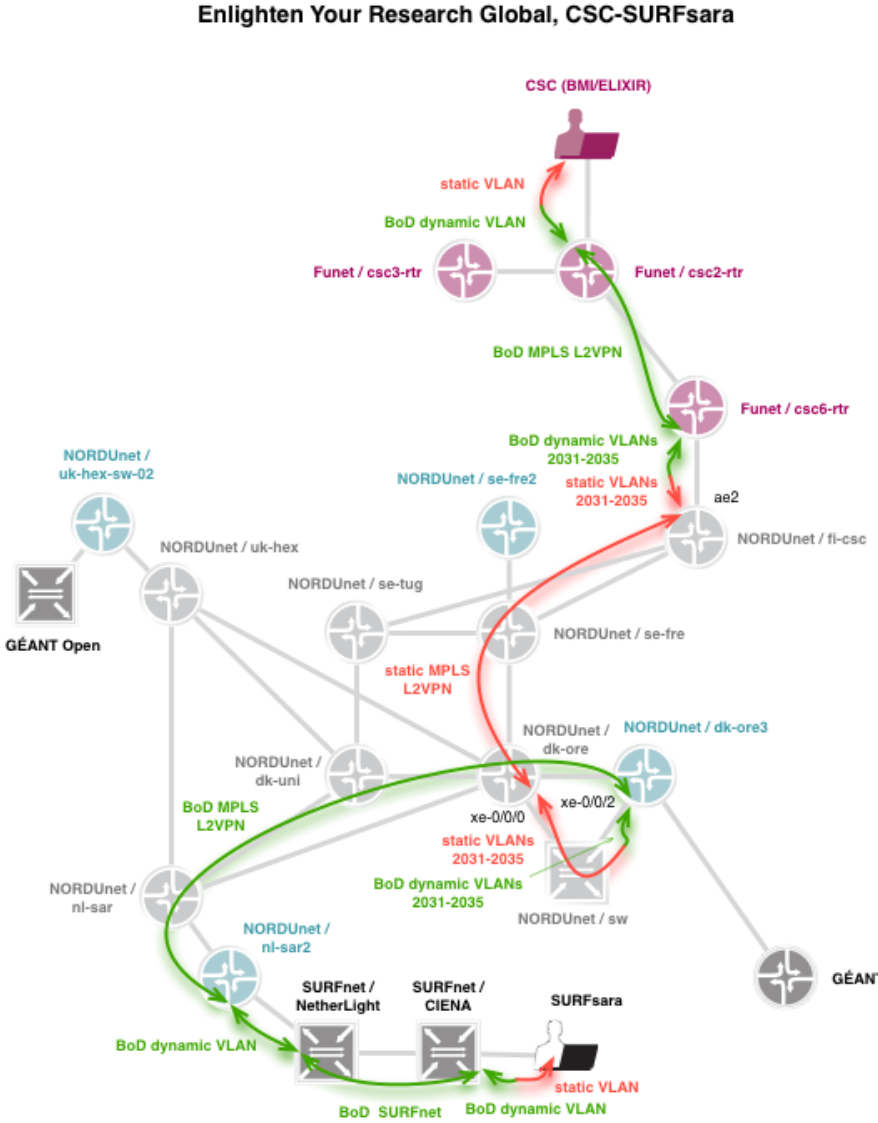


Collaborations with

- ELIXIR Competence Centre in EGI (FI, EMBL-EBI, CZ, FR, NL, GR)
  - Use cases to assess, use and adopt EGI cloud resources within the ELIXIR community
- Integrating ELIXIR reference datasets within the European Grid Infrastructure (IT, GR, Storage task force)
  - Coordinated effort to identify, and expose ELIXIR reference datasets within EGI
- EGI Federated Cloud adoption (Cloud task force)
  - Expose a common interface for existing ELIXIR resources
- Virtual Machine Library (Cloud task force)
  - Repository to upload ELIXIR VMI

Collaborations with GÉANT

- European ELIXIR Data - "LightPath" (EMBL-EBI, FI)
  - Explore the replication of large scale (Petabyte scale) archives to remote sites
- REMS - Resource Entitlement Management System (FI, EMBL-EBI)
  - Access to sensitive data granted by a Data Access Committee
- Cross-site VM Operation - EYRG (EMBL-EBI, FI, NL)
  - Transfer VMs between computing centers to allow researchers to perform analyses in the cloud
- AAI code of conduct
  - ELIXIR endorsement on a common way towards user attribute release in sync with eduGAIN.



Thanks!

Members

Observers

Potential Bottlenecks in Life Sciences

- Data production grows faster than storage
- Cost of data production technologies declines faster than storage
- It takes longer to transfer data than produce them

Potential solutions

- Storage
  - Data compression
  - Select what we store
    - Evaluate data reproducibility & value of data
- Network
  - Faster protocols
  - Partitioning
  - Network upgrade
- Computation
  - Clouds
  - Data close to computation

Factors that can influence data availability

- **scientific**: e.g. data reproducibility, uniqueness, value of processed and/or raw data
- **financial:** cost of data storage, transfer, reproduction
- **technical:** storage, network, computation…
- **political:** drivers e.g. from funding bodies/large organisations/national interests
- **social:** data sharing mentality of the community in question
- **legal/ethical/formal:** requirements/constraints for data storage/transfer/access - e.g. need to store data on German citizens in Germany; requirements from journal publishers, data management plans, etc.

Some conclusions

- Data growth will change how we do things today
- Opportunity for e-infrastructures to better understand BMS RI problems.
  - Identification of bottlenecks
  - Discussion of some potential solutions
- Different communities -> different models -> some common solutions
- Solutions have to come from use cases
- We need to use technology more efficiently
- BMS community has to evaluate the practicality of storing everything
- Privacy issues make big data more challenging
- Difficult to separate big data from computation
- BMS RI need to be better at defining their requirements
- Shortage of expertise of how to deal with scientific data and IT services