Proteomics repositories integration using EUDAT resources

Rafael C Jimenez

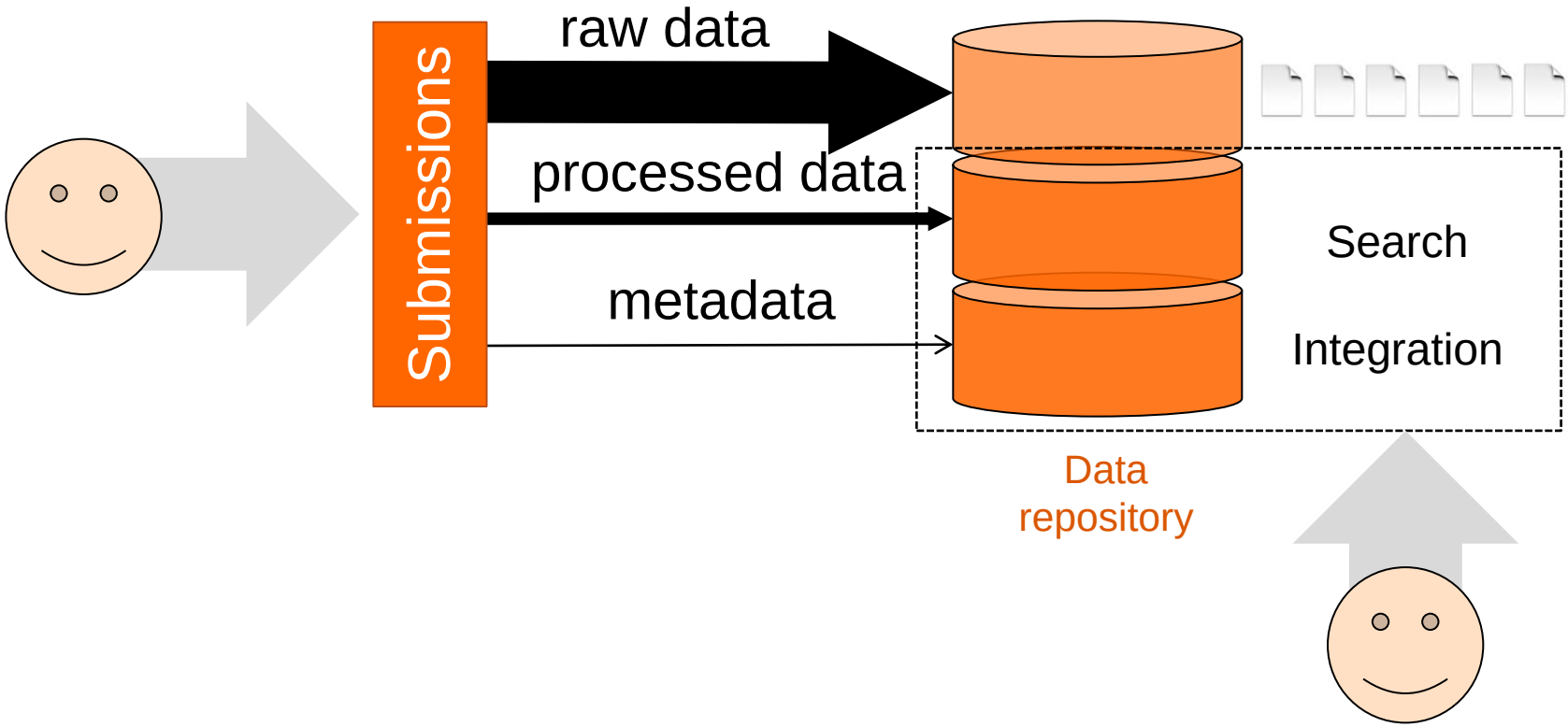
ELIXIR CTO

25 September 2014

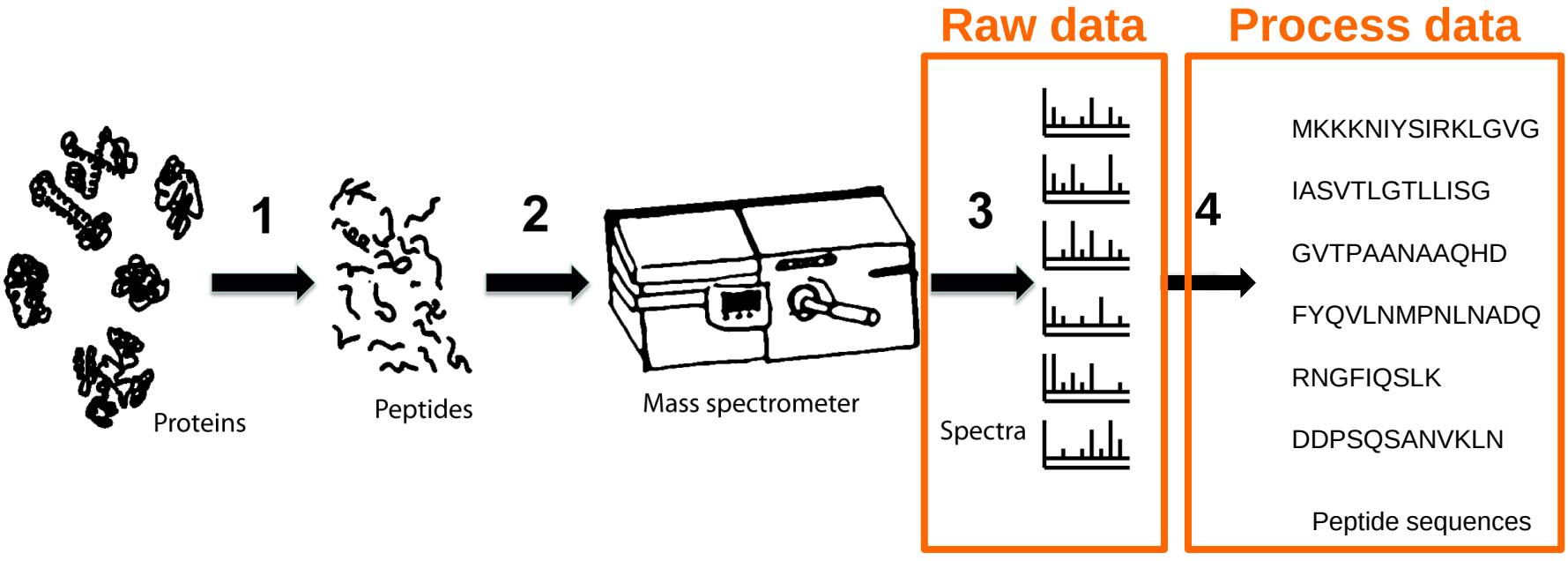
European Life Sciences Infrastructure for Biological Information

www.elixir-europe.org

Data submissions

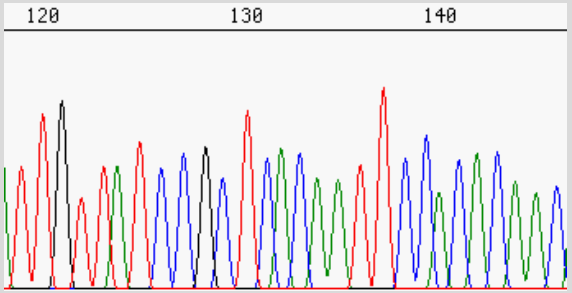
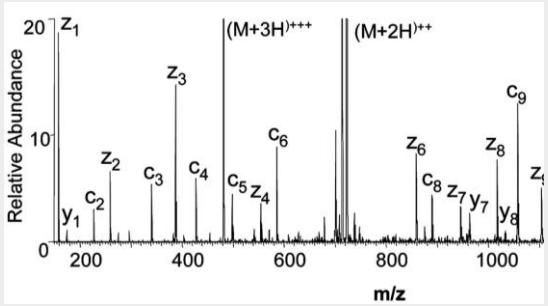


Overview of shotgun proteomics data production



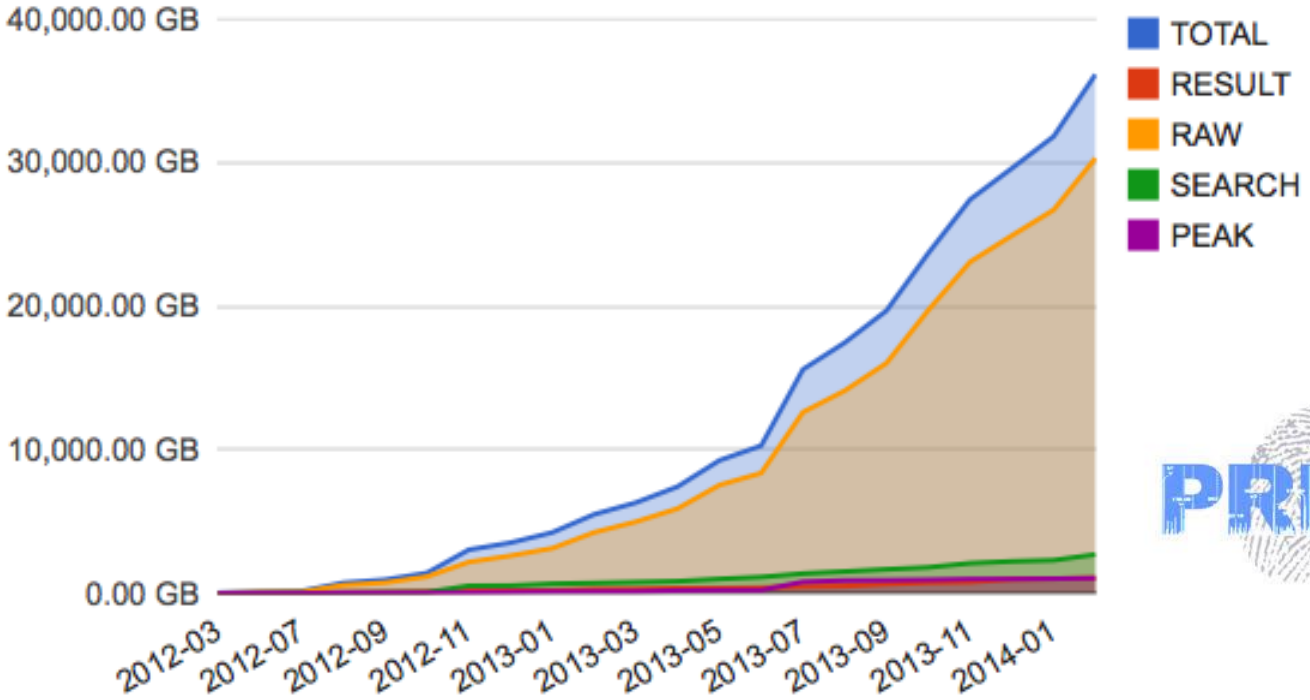
Noble WS, MacCoss MJ (2012) Computational and Statistical Analysis of Protein Mass Spectrometry Data. PLoS Comput Biol 8(1): e1002296. doi:10.1371/journal.pcbi.1002296
http://www.ploscompbiol.org/article/info:doi/10.1371/journal.pcbi.1002296

Data examples

Raw data	Process data	Metadata
	TTGTTATCCG…	***DNA*** *Human* *Liver* *Mitochondria* *W. Smith* …
	LPISASHSSK…	***Peptide*** *Mouse* *Heart* *Nucleus* *J. Heinz* …
…	…	…

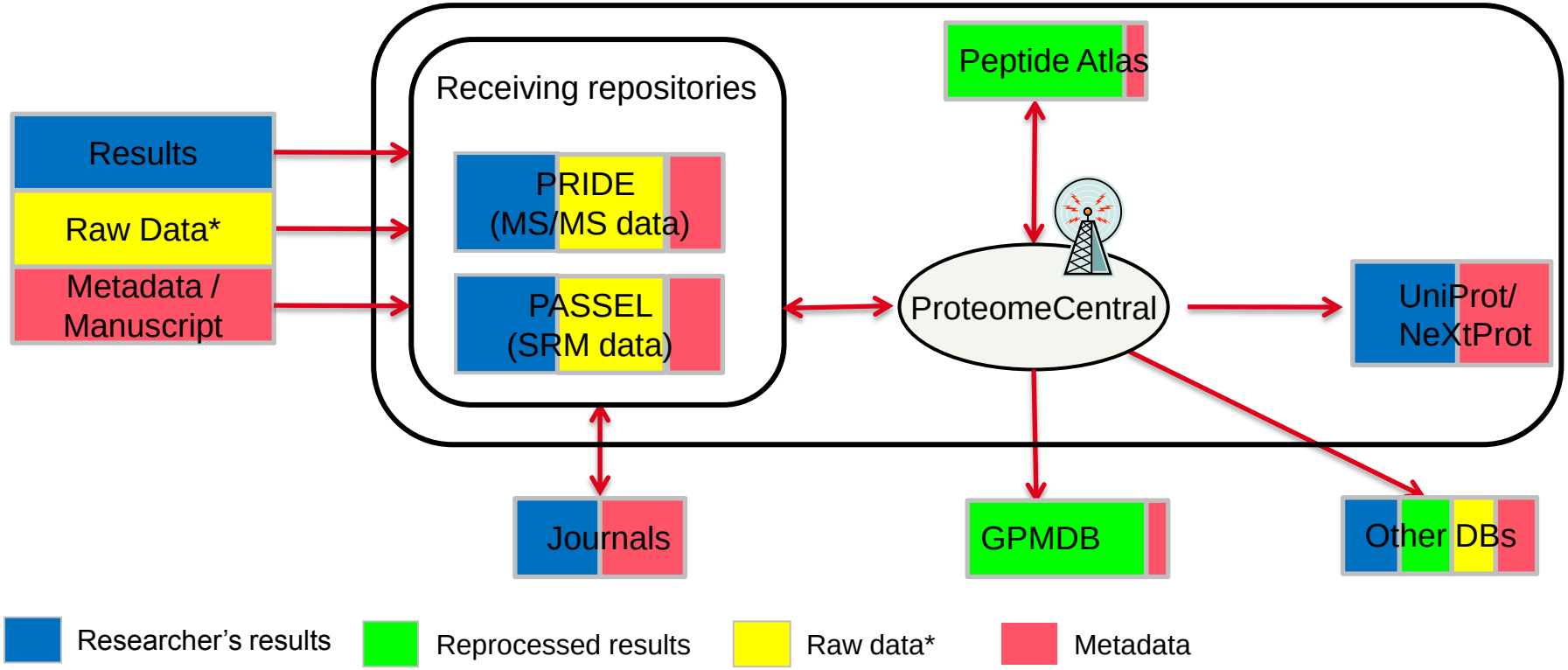
Proteomics data in PRIDE

~85% raw data



ProteomeXchange

- **Framework** to enable **standard data submission and dissemination pipelines** between the main existing proteomics resources.



Vizcaíno *et al.*, *Nature Biotechnology*, 2014

PRIDE (PRoteomics IDEntifications) database

EMBL-EBI

Services | Research | Training | About us

PRIDE Archive

Search

Examples: stress, human, blood

Home | Submit data | Browse data | Help | About PRIDE Archive | Register | Login | Feedback

PRIDE > Archive

PRIDE Archive - proteomics data repository

The PRIDE PRoteomics IDEntifications database is a centralized, standards compliant, public data repository for proteomics data, including protein and peptide identifications, post-translational modifications and supporting spectral evidence.

Datasets

- 1467 projects
- 31195 assays

Submit data

Submit data: we welcome direct submissions of MS/MS proteomics datasets.

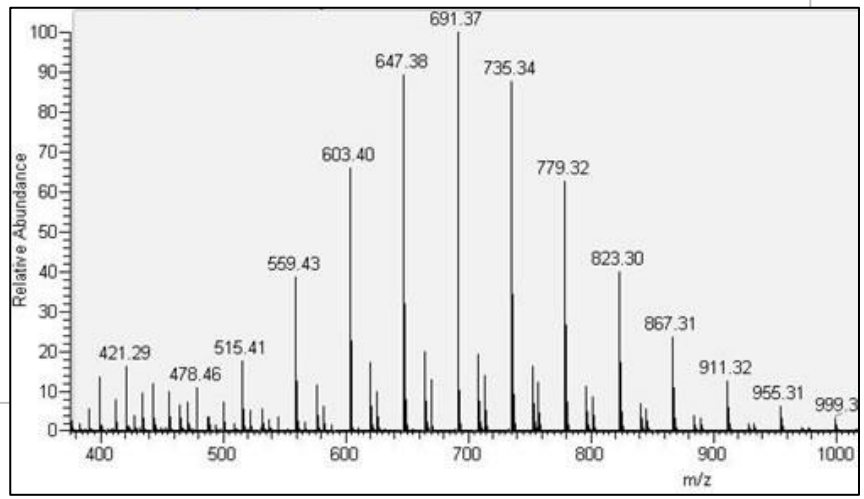
ProteomeXchange: accepts both MS/MS proteomics submissions and SRM submissions. PRIDE is the official entry point for MS/MS proteomics submissions.

Access data

Browse: explore all public datasets currently available in our repository.

BioMart: use PRIDE BioMart for custom queries, such as linking PRIDE identifications to REACTOME pathways

FTP: data can be downloaded directly from our FTP site.



Mass spectrometry

Martens *et al.*, Proteomics, 2005
Vizcaíno *et al.*, NAR, 2013

Origin:
152 USA
108 Germany
67 United Kingdom
53 Switzerland
48 Netherlands
42 China
42 Canada
41 France
36 Spain
33 Belgium
25 Australia
23 Sweden
17 Japan
16 Denmark
13 Norway
12 Finland
12 India
12 Taiwan
10 Italy
9 Republic of Korea
8 Austria
8 Ireland
8 Brazil
7 Singapore
5 Israel
5 Russia ...

Type:
273 PRIDE complete
501 PRIDE partial
47 PeptideAtlas/PASSEL complete

Submissions/year:
2012: 102
2013: 527
2014: 192

Access:
38.3% PRIDE public
5.3% PASSEL public
56% PRIDE private
0.4% PASSEL private

Top Species studied by at least 8 datasets:
381 *Homo sapiens*
100 *Mus musculus*
31 *Arabidopsis thaliana*
26 *Saccharomyces cerevisiae*
16 *Escherichia coli*
14 *Rattus norvegicus*
12 *Mycobacterium tuberculosis*
11 *Drosophila melanogaster*

~ 215 species in total

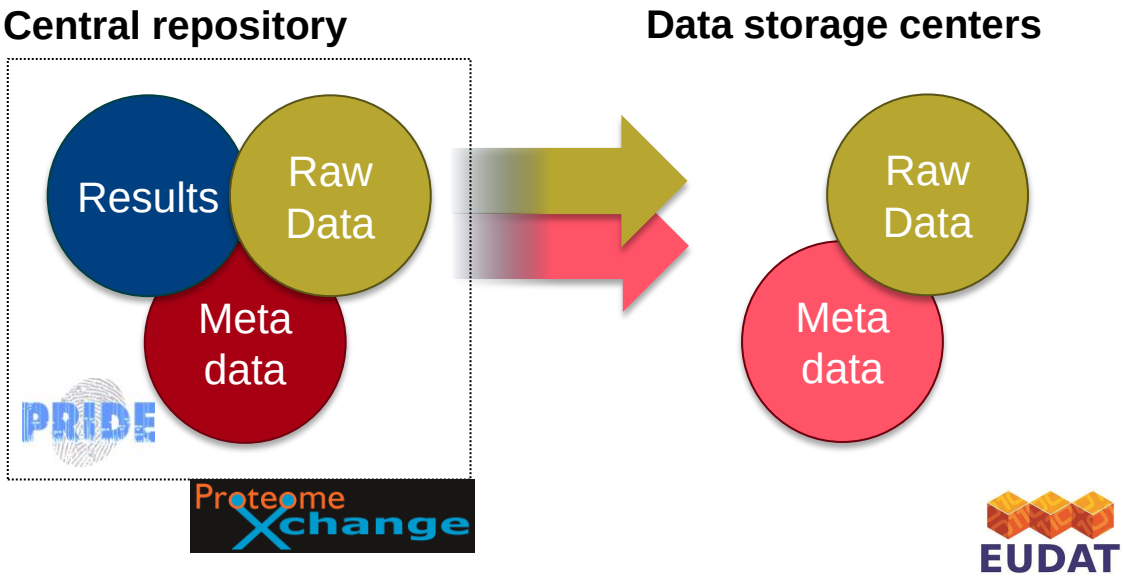
Data volume:
Total: **>40 TB**
Number of all files: >120,000
PXD000320-324: ~ 5 TB
PXD000065: ~ 1.4TB

Pilot evolution

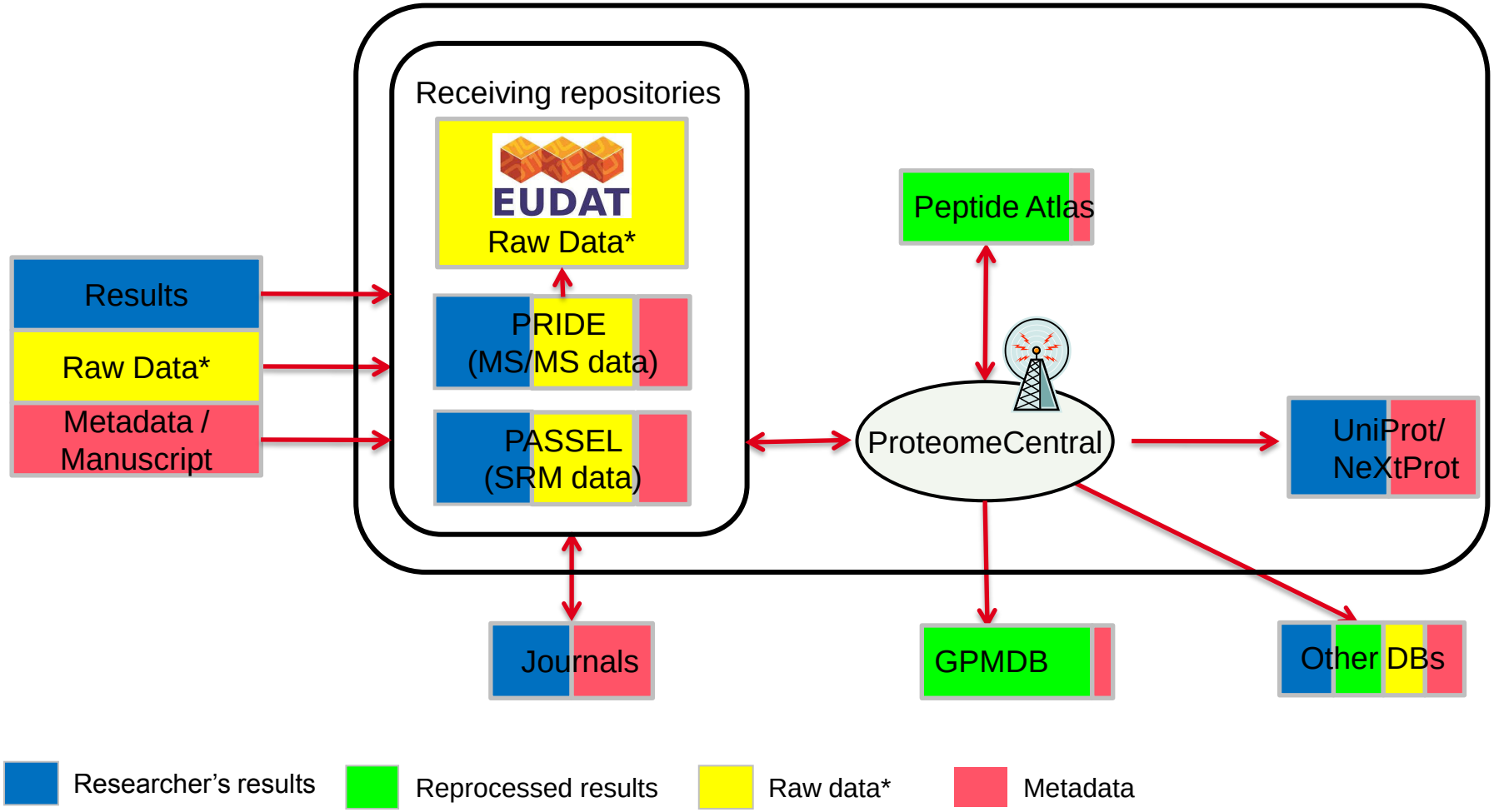
- Use EUDAT
  - Replication of ELIXIR data in EUDAT data centers
  - Delegation of ELIXIR data in EUDAT data centers

- Adopt EDUAT
  - Replication of ELIXIR data in ELIXIR data centers using EUDAT technology

Replication of ELIXIR data in EUDAT data centers

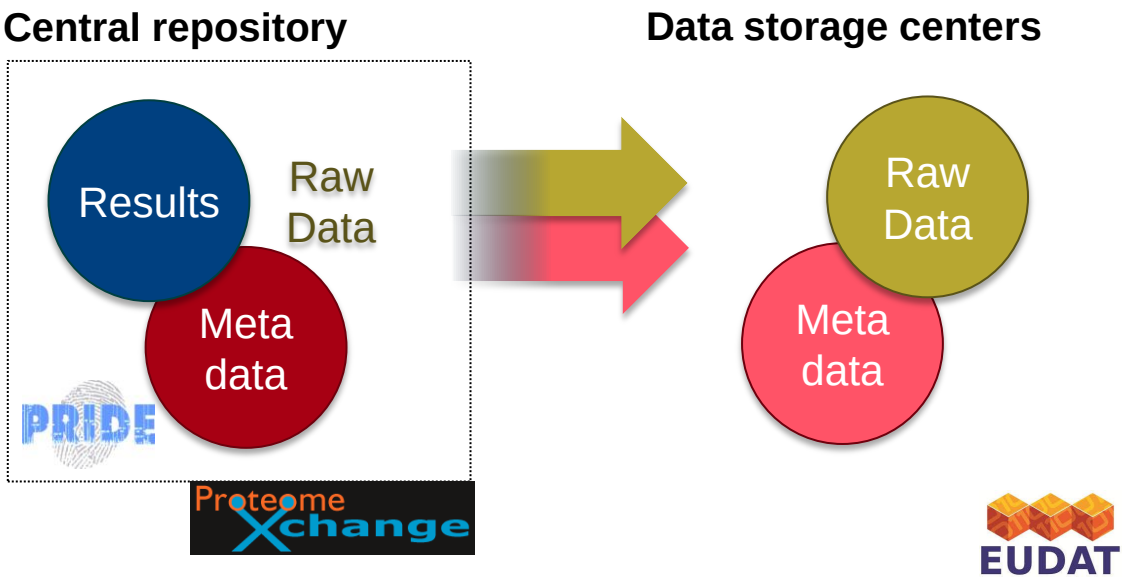


Replication of ELIXIR data in EUDAT data centers

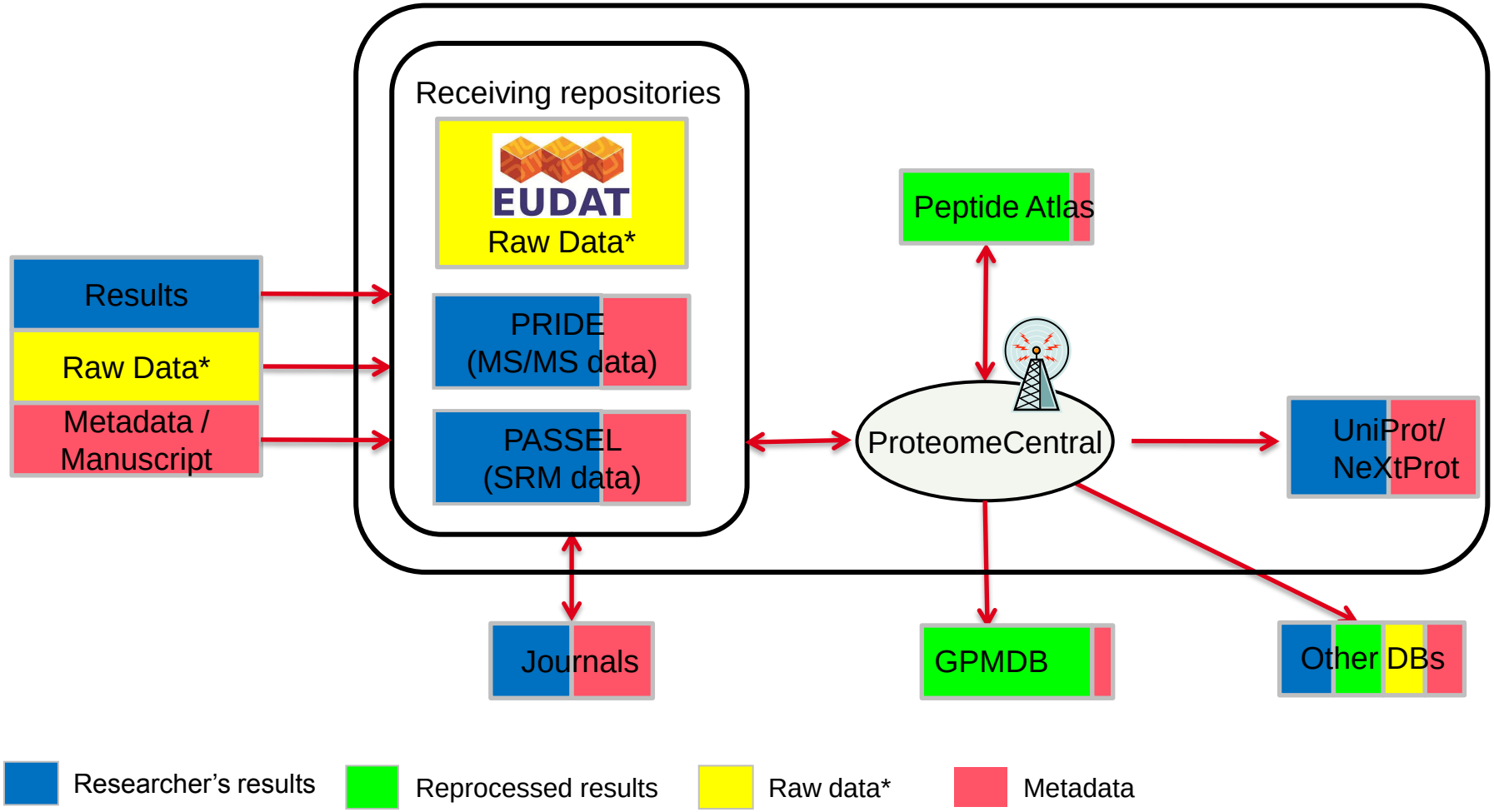


Vizcaíno *et al.*, *Nature Biotechnology*, 2014

Delegation of ELIXIR data in EUDAT data centers

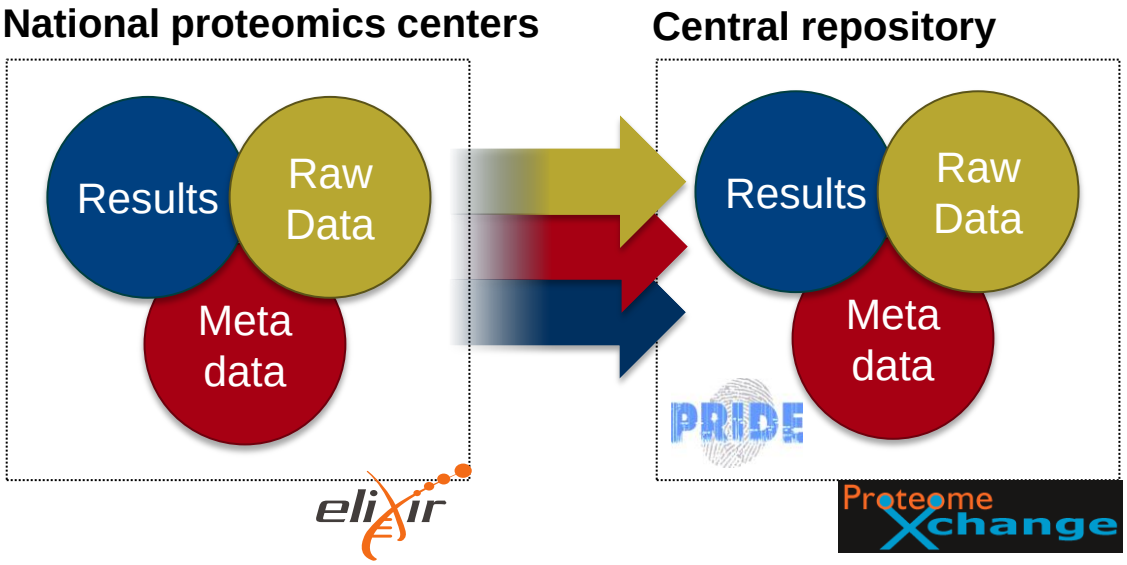


Delegation of ELIXIR data in EUDAT data centers



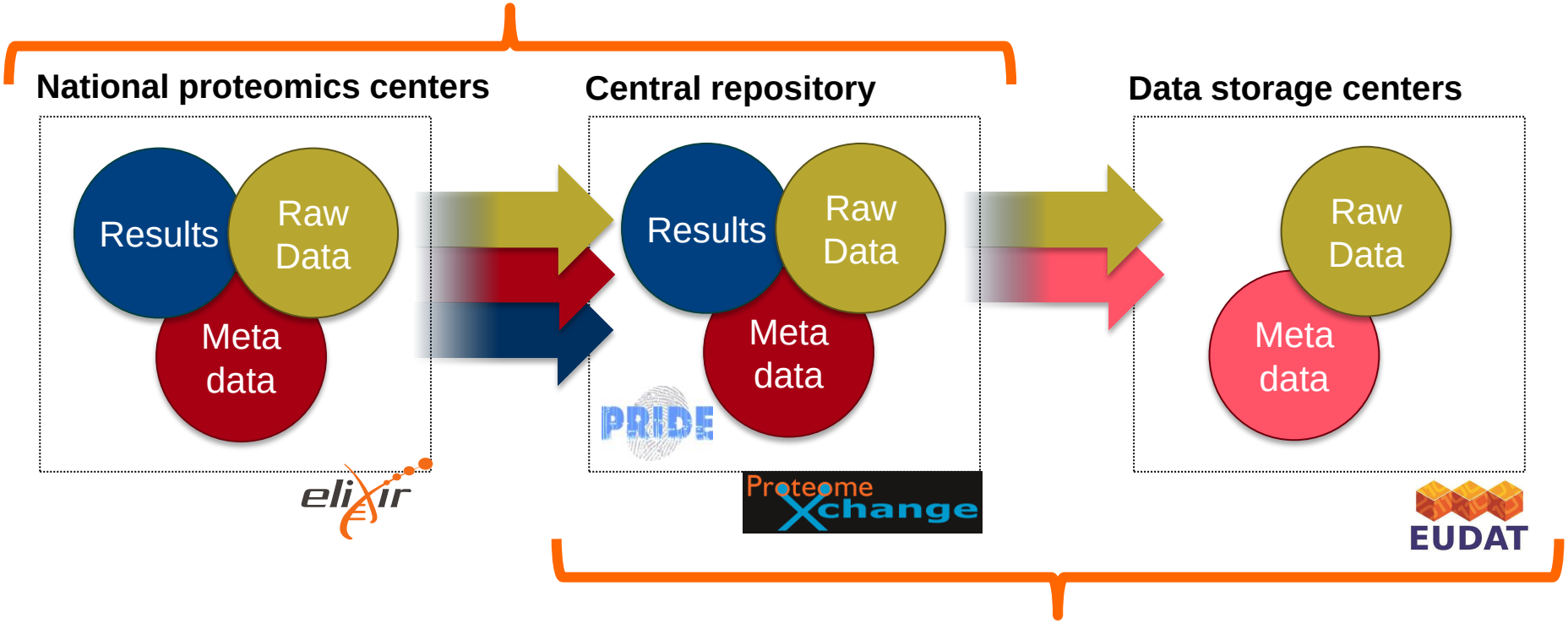
Vizcaíno *et al.*, *Nature Biotechnology*, 2014

Replication of ELIXIR data in ELIXIR data centers using EUDAT technology



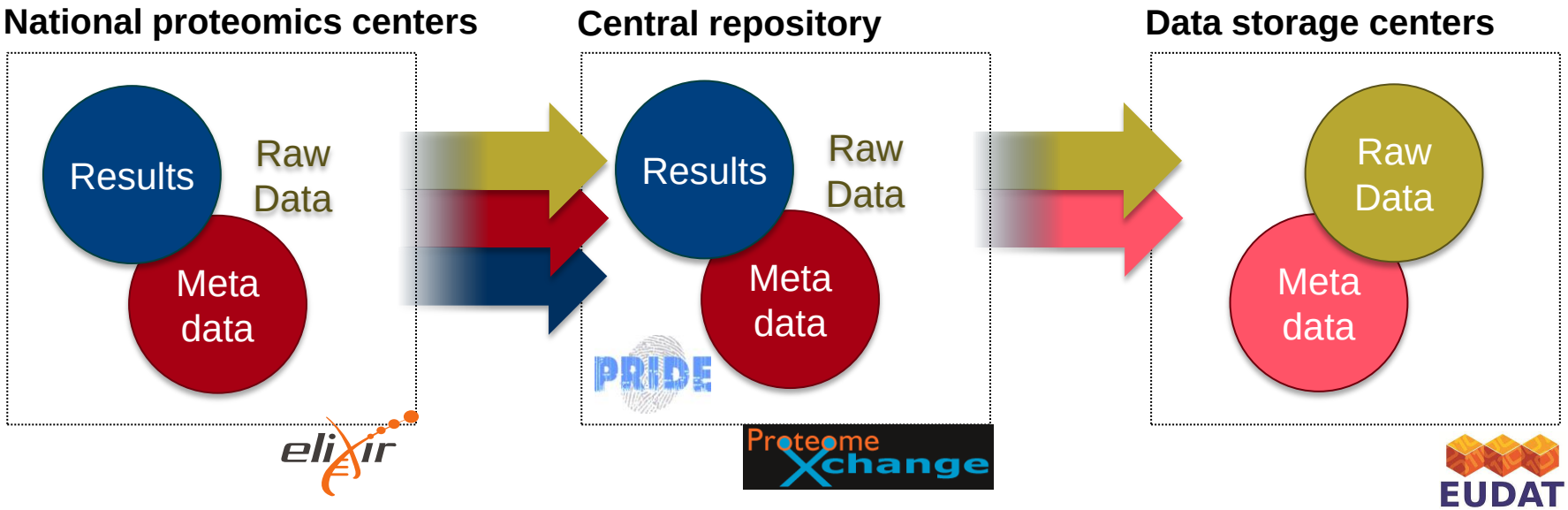
Plans

1.- ELXIR replication



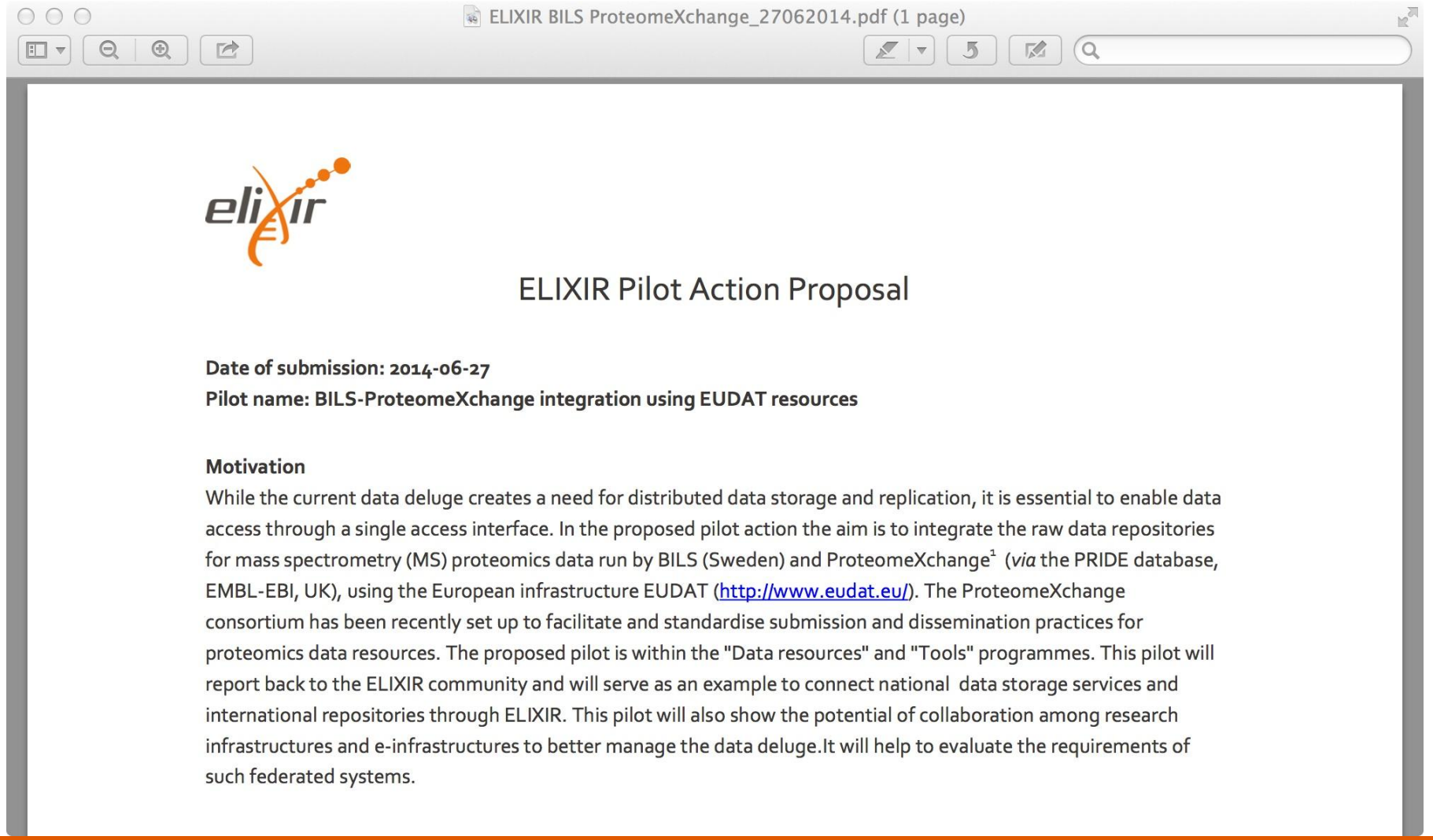
2.- EUDAT replication

Plans



3.- delegation

ELIXIR Pilot action


ELIXIR BILS ProteomeXchange_27062014.pdf (1 page)

ELIXIR Pilot Action Proposal

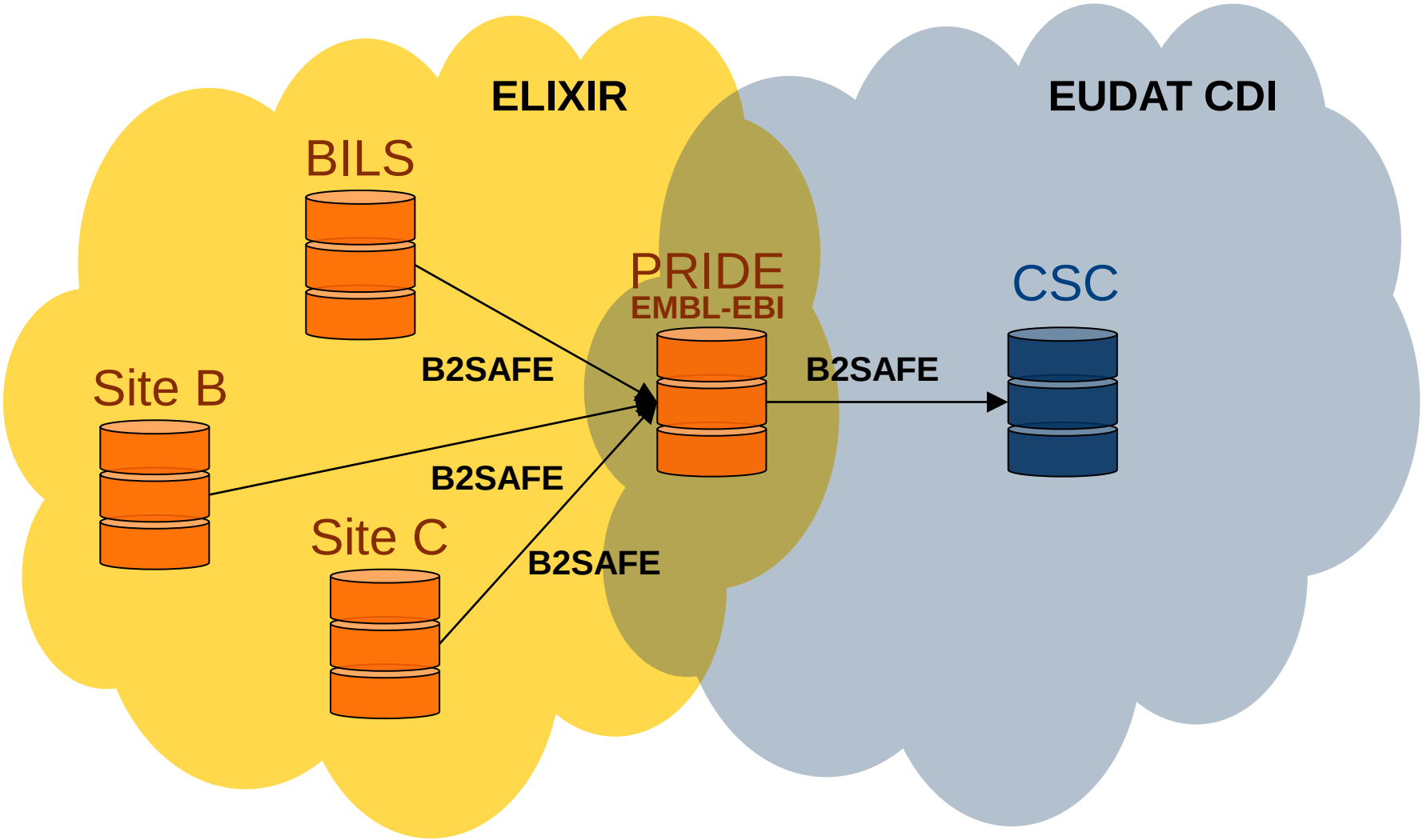
Date of submission: 2014-06-27
Pilot name: BILS-ProteomeXchange integration using EUDAT resources

Motivation

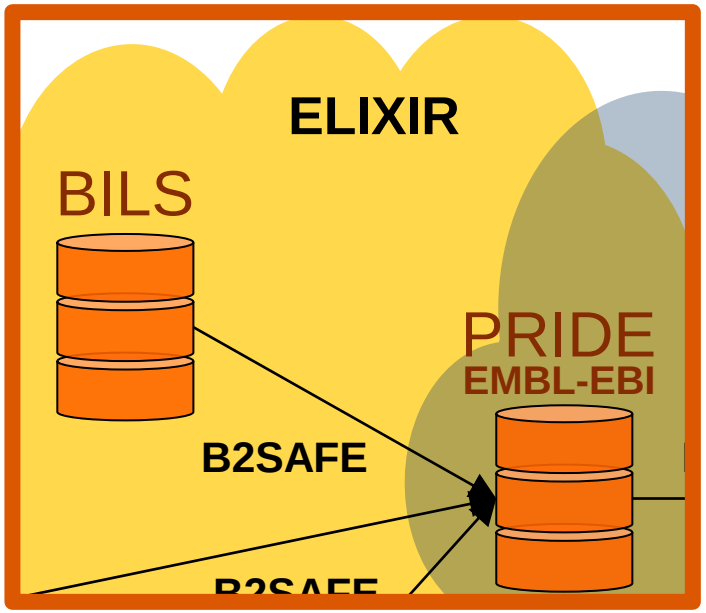
While the current data deluge creates a need for distributed data storage and replication, it is essential to enable data access through a single access interface. In the proposed pilot action the aim is to integrate the raw data repositories for mass spectrometry (MS) proteomics data run by BILS (Sweden) and ProteomeXchange[1] (*via* the PRIDE database, EMBL-EBI, UK), using the European infrastructure EUDAT (http://www.eudat.eu/). The ProteomeXchange consortium has been recently set up to facilitate and standardise submission and dissemination practices for proteomics data resources. The proposed pilot is within the "Data resources" and "Tools" programmes. This pilot will report back to the ELIXIR community and will serve as an example to connect national data storage services and international repositories through ELIXIR. This pilot will also show the potential of collaboration among research infrastructures and e-infrastructures to better manage the data deluge.It will help to evaluate the requirements of such federated systems.

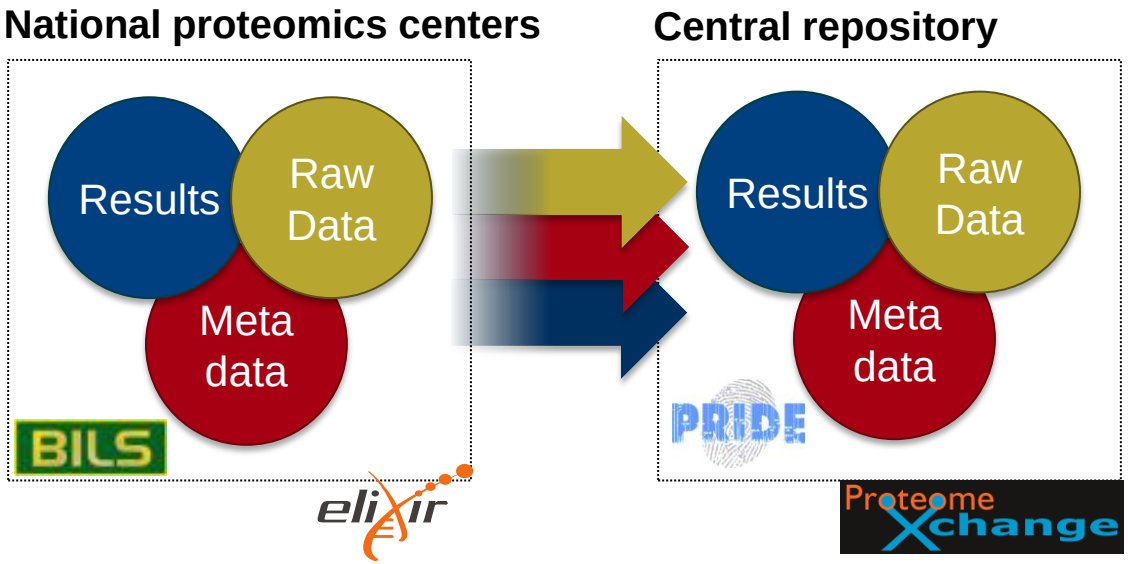
EUDAT services

File sharing model

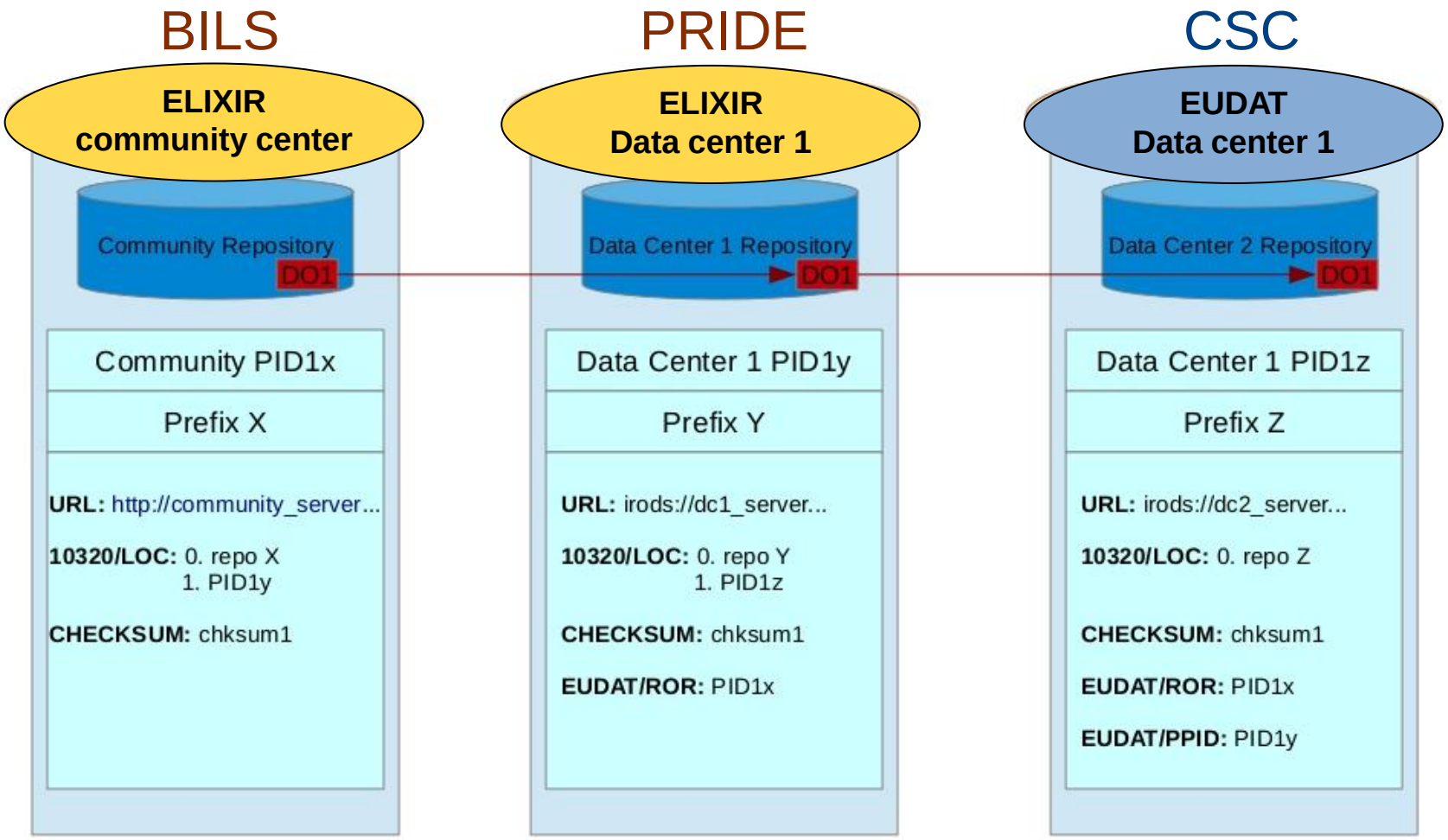


Pilot – EUDAT adoption: ELIXIR replication





PIDs


BILS
ELIXIR
community center
Community Repository
DO1
Community PID1x
Prefix X
URL: http://community_server...
10320/LOC: 0. repo X
1. PID1y
CHECKSUM: chksum1
PRIDE
ELIXIR
Data center 1
Data Center 1 Repository
DO1
Data Center 1 PID1y
Prefix Y
URL: irods://dc1_server...
10320/LOC: 0. repo Y
1. PID1z
CHECKSUM: chksum1
EUDAT/ROR: PID1x
CSC
EUDAT
Data center 1
Data Center 2 Repository
DO1
Data Center 1 PID1z
Prefix Z
URL: irods://dc2_server...
10320/LOC: 0. repo Z
CHECKSUM: chksum1
EUDAT/ROR: PID1x
EUDAT/PPID: PID1y

Status

- BILS
  - Migrating from existing Swestore dCache to iRODS
  - Testing compatibility with B2SAFE
    - Latest iRDOS not compatible with B2SAFE?

- PRIDE
  - iRODS service installed
  - B2SAFE module have been deployed at EMBL-EBI (PRIDE)
  - Test B2SAFE replication PRIDE -> CSC
  - DOI for datasets
  - PID for dataset files
  - Web service to associate datasets to dataset files

Status

In progress

- Handle System Registration
- Test requests of EPIC/EUDAT identifiers

Open questions

- BILS local PIDs?
- Sync back from PRIDE to BILS for modifications/additions at PRIDE?
- Data push or pull model?
- Replication of process data requires previous validation

Participants

EUDAT/CSC

- Jani Heikkinen
- Damien Lecarpentier

EMBL-EBI/systems

- Andy Jenkinson
- Steven Newhouse

EMBL-EBI/PRIDE

- Juan Antonio Vizcaíno
- Henning Hermjakob

BILS

- Mikael Borg
- Fredrik Levander
- Bengt Persson

ELIXIR Hub

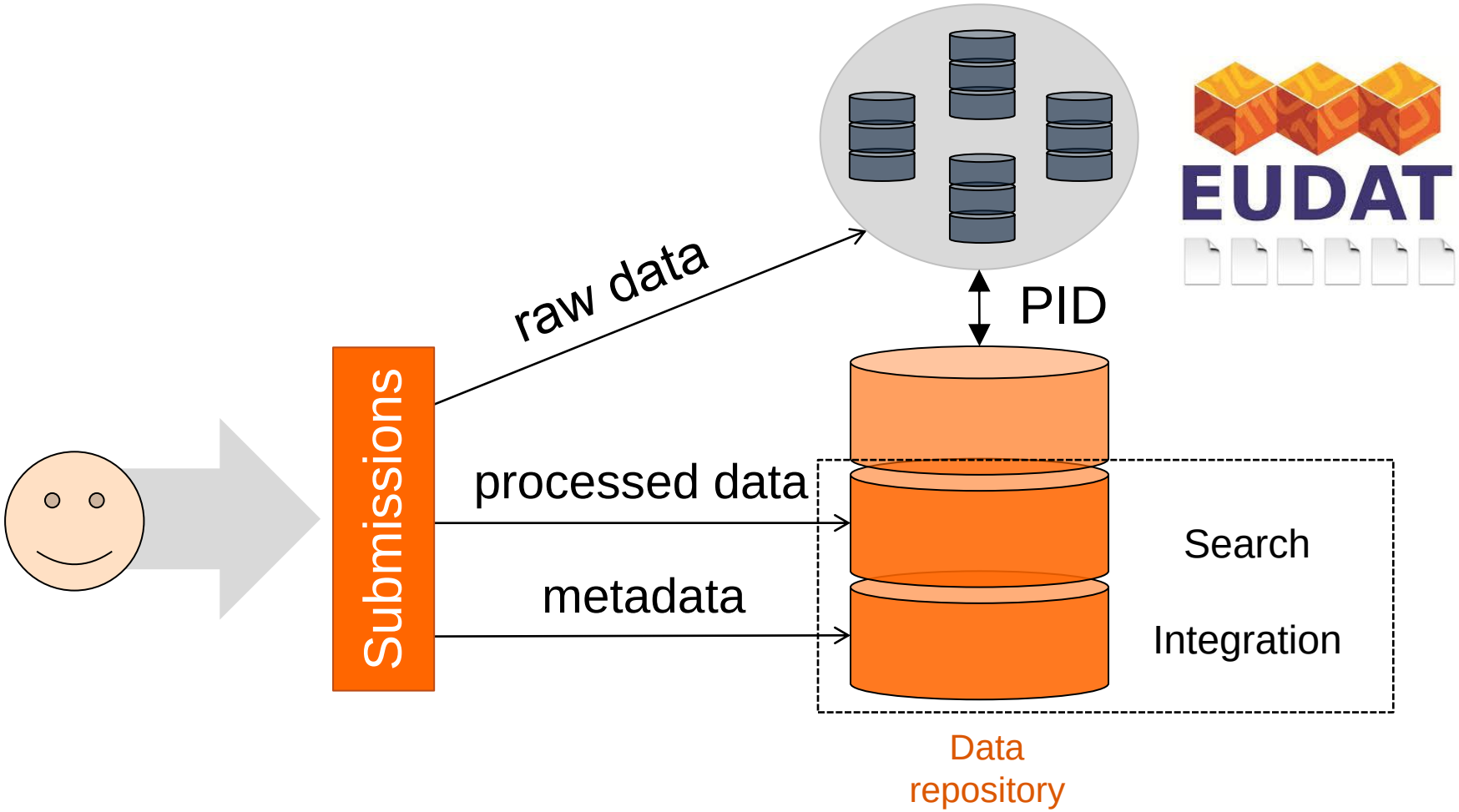
- Rafael C Jimenez

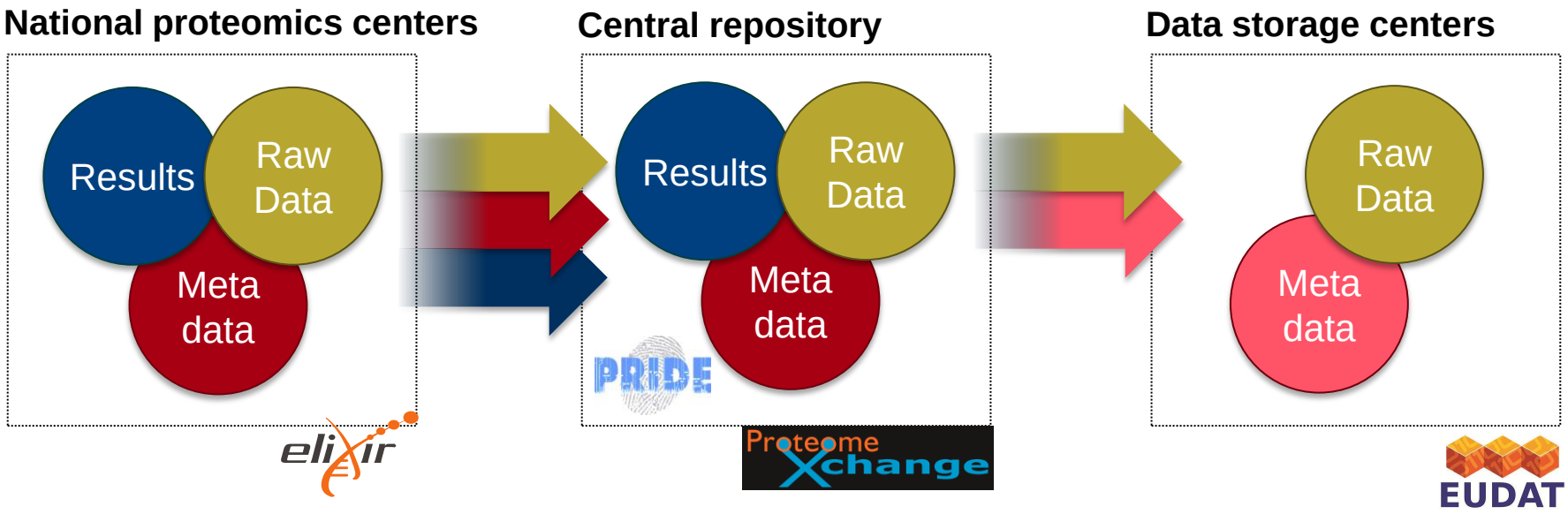
Thank you for your attention

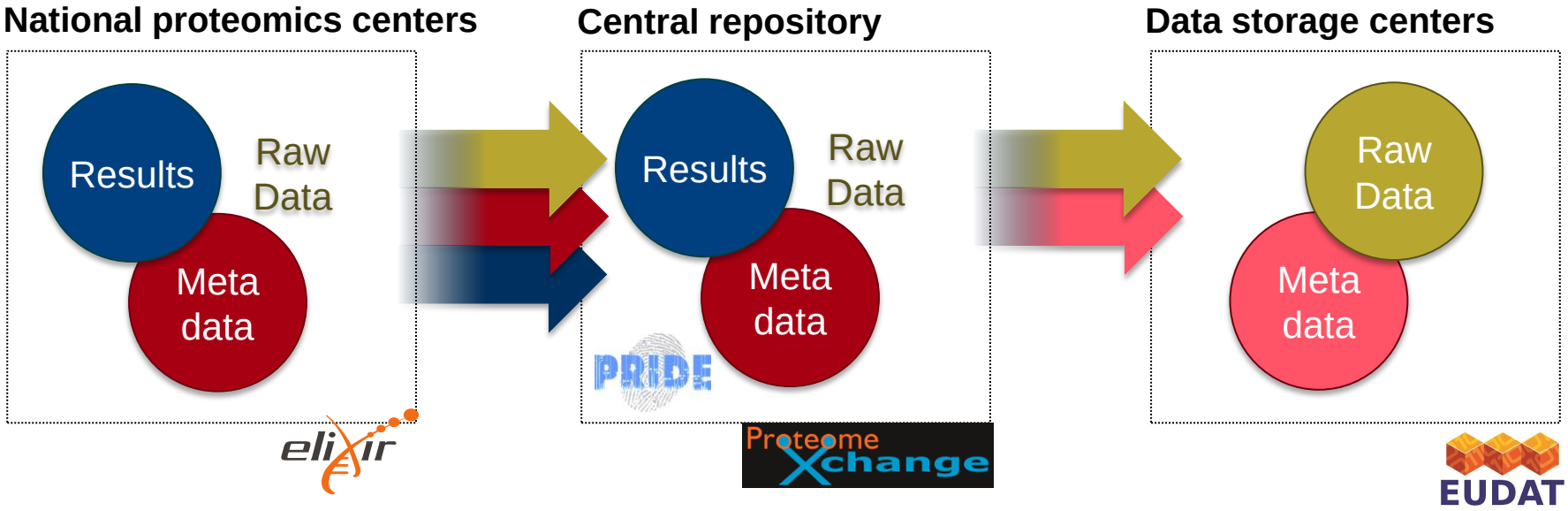
European Life Sciences Infrastructure for Biological Information

www.elixir-europe.org

Delegation of raw data




National proteomics centers
Central repository
Data storage centers
Results
Raw Data
Meta data
PRIDE
ProteomeXchange
elixir
EUDAT


National proteomics centers
Central repository
Data storage centers
Results
Raw Data
Meta data
PRIDE
ProteomeXchange
elixir
EUDAT