Data management and data harmonisation for large collaborative biobank studies

Maria Krestyaninova

IT for collaborative innovation

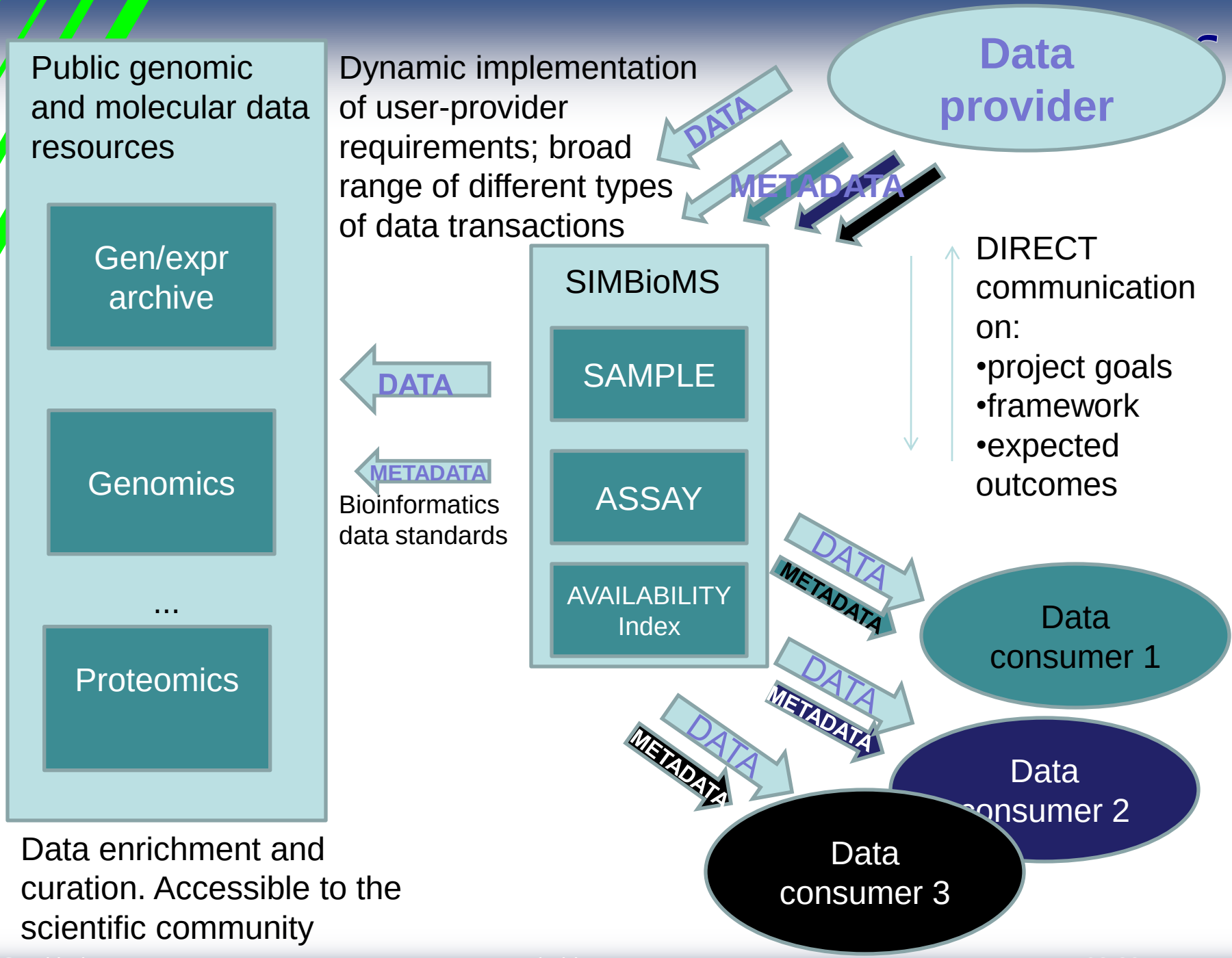
highly varied level and type of in-house data management solutions

public infrastructure geared towards amassing data rather than exchange

need for dynamic and flexible information exchange environment to facilitate dialoguing, and tracking IP ownership and to speed up the discovery

Biobank IT use-cases

•Sample and assay tracking, house keeping data management
Internal biomaterial management: collection, distribution and QA
Users: local biobank maintenance and distribution team
Example: OBIBA, any LIMS solution
Data analysis, internal
•Data access for analysts: advanced search, analytical and vizualisation tools
Users: statisticians and epidemiologists
Example: Bioconductor
Harmonisation, annotation and indexing of data
•Data enrichment and annotation; mapping to standard ontologies and vocabularies
Users: data managers and epidemiologists world-wide
Example: DataSHaPER, SAIL, etc
•Controlled data and/or metadata release for collaboration enhancement
Selective and secure data sharing with collaborators and scientific community
Users: potential and existing collaborators, external
Example: EGA, SAIL, SIMS/AIMS, ISATAB


Public genomic and molecular data resources
Gen/expr archive
Genomics
...
Proteomics
Dynamic implementation of user-provider requirements; broad range of different types of data transactions
Data provider
DATA
METADATA
SIMBioMS
SAMPLE
ASSAY
AVAILABILITY Index
DATA
METADATA
Bioinformatics data standards
DIRECT communication on:
•project goals
•framework
•expected outcomes
DATA
METADATA
Data consumer 1
DATA
METADATA
Data consumer 2
DATA
METADATA
Data consumer 3
Data enrichment and curation. Accessible to the scientific community

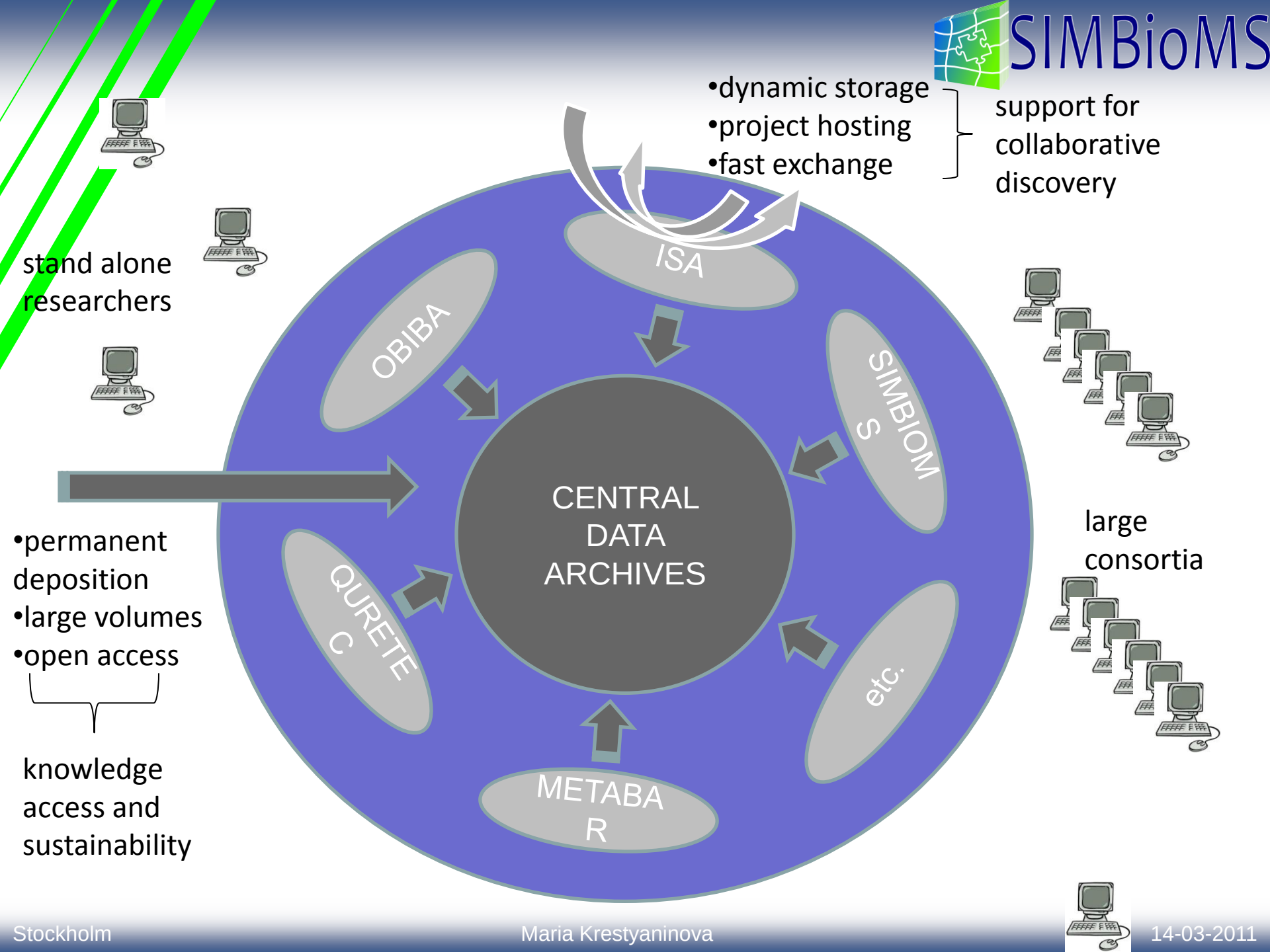
International initiatives aiming to enhance biobank interconnectivity

- Public Population Project in Genomics (P3G)
  - www.p3g.org
- Biobanking and Biomolecular Resources Research Infrastructure (BBMRI)
  - www.bbmri.eu

Areas of interest for computer scientists

- metadata
- semantic enrichment
- standards

Data integration and data management solutions


SIMBioMS
•dynamic storage
•project hosting
•fast exchange
support for collaborative discovery
stand alone researchers
ISA
OBIBA
SIMBIOMS
CENTRAL DATA ARCHIVES
QURETEC
etc.
METABAR
large consortia
•permanent deposition
•large volumes
•open access
knowledge access and sustainability

- www.obiba.org – open source s/w for biobanks

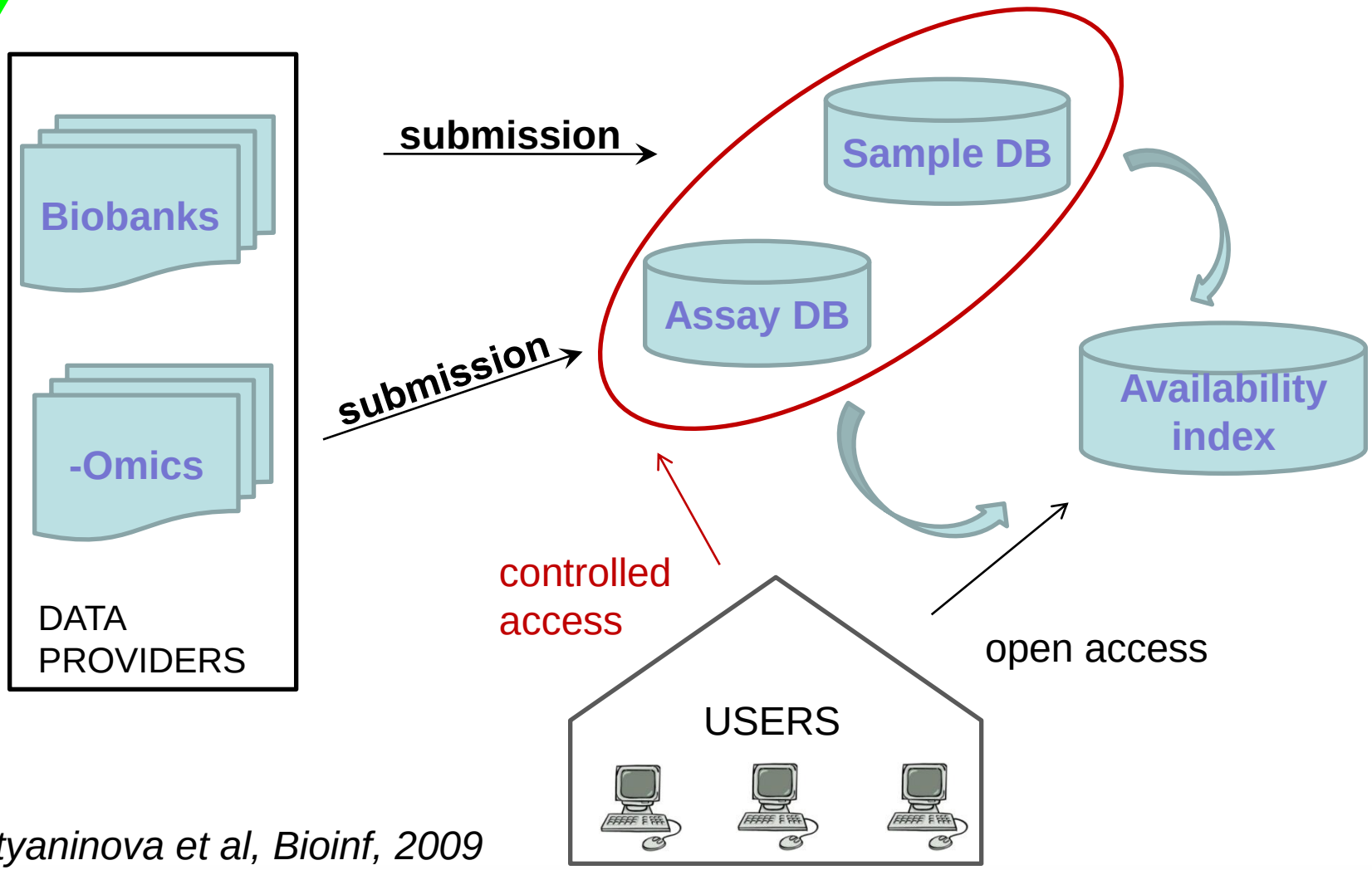
- www.quretec.com – s/w platform for data management, collection and analysis in clinical research

- http://isatab.sourceforge.net/ – ISA (Investigation/Study/Assay) Infrastructure

- http://simbioms.org – open source solution for large-scale collaborative projects

Banet et al, BMC Bioinf, submitted

System overview



Krestyaninova et al, Bioinf, 2009

Emäntä

MISSION:

Assist researchers and administrators in **reporting and project management**

FUNCTIONALITY:

Initiating new projects and deadline alerts

Activity notifications (who's uploading data, when)

Transparency of publications *vs* contribution

support@simbioms.org

Assay and Sample Information Management modules

MISSION:

Central secure data **storage** and **annotation**, standard compliance

FUNCTIONALITY:

Fast upload/download

Easy metadata capture

Export to public archives

Connectivity with analytical pipelines and other resources

Collaborations with other consortia; selective data sharing

Harmonisation

Central indexing of data stored locally

- Heavy or protected data remains in original location
- Availability of data points or samples is indexed centrally against Variables of Interest
- Amount and location of data relevant to a research question is estimated online

http://sail.simbioms.org

Currently

-3 production instances

-~200 000 samples characterised with more than 100 variables of interest

->20 data contributors to the common data index of biomaterials across-europe

- collaboration with biospecimen catalogues and ontology initiatives

Near future

-tighter integration with suppliers of standard ontologies and vocabularies

- building on longitudinal data handling expertise in other domains

-wider data contributor circle through interaction with national biobanking initiatives

Gostev et al, Bioinf, 2010

WELCOME TO DATASHAPER WEBSITE

The DataSHaPER (Data Schema and Harmonization Platform for Epidemiological Research) is both a scientific approach and a suite of practical tools. Its primary air facilitate the prospective harmonization of emerging biobanks, provide a template for retrospective synthesis and support the development of questionnaires and inform collection devices, even when pooling of data with other biobanks is not foreseen.

- **What is the DataSHaPER?**

The development of these tools has been jointly funded by **P³G**, **CPT**, **PHOEBE**, and **Generation Scotland**.

DataSchema Platform

A DataSchema identifies and describes a thematic set of core variables that are of particular value in a specified sci setting.
It also contains associated support material including variable definitions, links to relevant Ontologies and Classifica access to reference questionnaires and operating procedures.

Harmonization Platform

The Harmonization Platform provides a template for the formal estimation of the potential to synthetized information studies. At present, access to the Harmonization Platform is limited to collaborative context.

Please be aware of the **terms and conditions** before proceeding.

Current infrastructural volume

12 installation in 4 countries

>100 user-organisations

>200.000 samples

>50.000 assays and studies

9 international research and infrastructure projects

Krestyaninova et al, ***Bioinformatics***, 2009
Viksna et al, ***BMC Bioinformatics***, 2007
Banet at al, ***BMC Bioinformatics***, submitted

Inter-institutional team

Uniquer

- Russell Vincent
- Joern Dietrich

EMBL-EBI

- Stathis Kanterakis
- Ugis Sarkans

Institute of Mathematics and Computer Science
University of Latvia

- Juris Viksna
- Sandra Ose
- Martins Opmanis
- Andris Zarins

FIMM
Institute for Molecular Medicine Finland
Nordic EMBL Partnership for Molecular Medicine

- Teemu Perheentupa
- Jani Heikkinen
- Samuli Ripatti
- Huei-Yi Shen