

# Data management and data harmonisation for large collaborative biobank studies

Maria Krestyaninova

FIMM, Finland

Cambridge University, UK

Uniquer, Switzerland



Institute of Mathematics and Computer Science  
University of Latvia



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

EMBL-EBI



**Uniquer**

- Juris Viksna
- Sandra Ose
- Martins Opmanis
- Andris Zarins
  
- Teemu Perheentupa
- Jani Heikkinen
- Samuli Ripatti
  
- Julio Fernandez Banet
- Ugis Sarkans
- Joern Dietrich
- Stathis Kanterakis
  
- Russell Vincent

# 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

# International initiatives aiming to enhance biobank interconnectivity

- Public Population Project in Genomics (P3G)

- [www.p3g.org](http://www.p3g.org)

- Biobanking and Biomolecular Resources Research Infrastructure (BBMRI)

- [www.bbmri.eu](http://www.bbmri.eu)

ABOUT P<sup>3</sup>G

- » [At a Glance](#)
- » [Organization](#)
- » [Charter of Principles](#)
- » [Board](#)
- » [Annual Report '08](#)
- » [FAQ](#)

## MEMBERSHIP

- » [Membership Conditions](#)
- » [Charter Members](#)
- » [Associate Members](#)
- » [Individual Members](#)
- » [Become a Member](#)

P<sup>3</sup>G EVENTS

- » **CHANGE OF VENUE**  
P<sup>3</sup>G Annual Meeting  
April 26-27, 2010  
**Montreal**
- » P<sup>3</sup>G General Meeting  
Sept 29-30, 2009  
Luxembourg

## WELCOME

The Public Population Project in Genomics (P<sup>3</sup>G) is a not-for-profit international consortium that provides the international population genomics community with easy access to the expertise, resources, innovative tools and most up-to-date information from all areas of public population genomics.

P<sup>3</sup>G works with biobankers and other subject-matter experts from around the world to:

- Encourage** collaboration between researchers and biobankers
- Promote** harmonization of information
- Optimize** the design, set-up and research activities of population-based biobanks
- Facilitate** the transfer of knowledge and provide training to those working in the field

P<sup>3</sup>G brings the genomics community closer together in person - at conferences and meetings held around the world, and online through the Observatory - a publicly accessible knowledge database. Principles of transparency and collaboration are integral to the P<sup>3</sup>G approach.

[Click to Become a Member](#)

NEWS & P<sup>3</sup>G EVENTS

- » **CHANGE OF VENUE**  
P<sup>3</sup>G Annual Meeting  
April 26-27, 2010  
**Montreal**

## COMMUNICATIONS

- » [e-update](#)  
January 2010
- » [Newsletter](#)  
September 2009
- » [P<sup>3</sup>G e-Brochure](#)

## SPONSORS





**BBMRI**  
Biobanking and  
Biomolecular  
Resources Research  
Infrastructure

## Managing resources for the future of biomedical research

[Home](#)

[News & Events](#)

[Proposals & Calls](#)

[About BBMRI](#)

[The Mission](#)

[The Facility](#)

[Background](#)

[About Biobanks](#)

[Workpackages](#)

[Partners & Membership](#)

[Publications & Reports](#)

[Stakeholder's Forum](#)

[Press](#)

[FAQ](#)

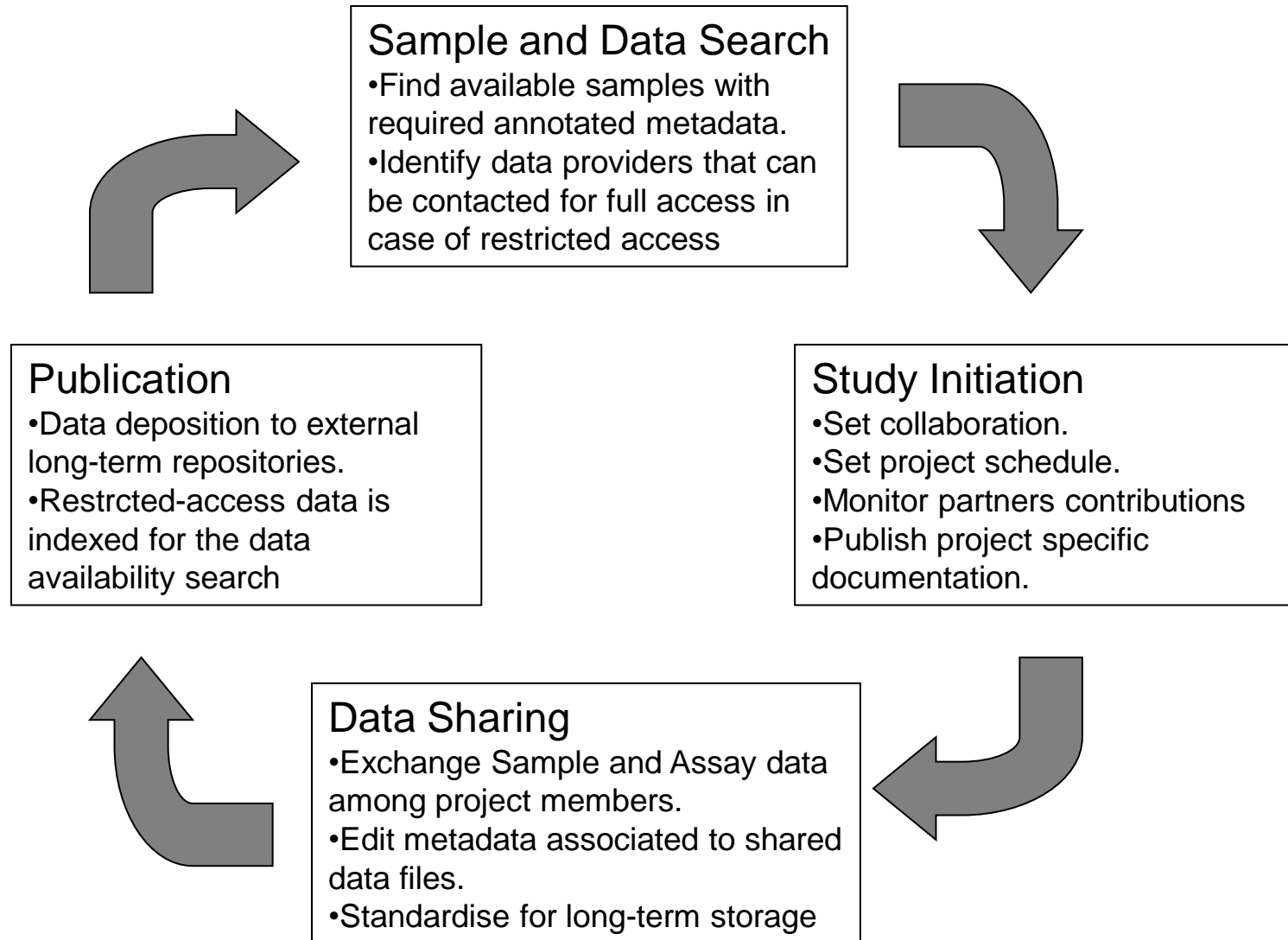
[Links](#)

### About Biobanks

This project aims at building a coordinated, large scale European infrastructure of biomedically relevant, quality-assessed mostly already collected samples (with the possibility to link to related clinical and epidemiological information), to enhance therapy and prevention of common and rare diseases, including cancer. In this area of unique European strength, valuable and irreplaceable national collections typically suffer from underutilisation due to fragmentation. Major synergism, gain of statistical power and economy of scale will be achieved by interlinking, standardising and harmonising - sometimes even just cross-referencing - a large variety of well-qualified, up-to date, existing and de novo national resources. The network should cover (1) major European biobanks with blood, serum, tissue or other biological samples, (2) molecular methods resource centres for human and model organisms of biomedical relevance, (3) and biocomputing centres to ensure that databases of samples in the repositories are dynamically linked to existing databases and to scientific literature as well as to statistical expertise.



# Use-cases



# 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 visualization 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

Data enrichment and curation. Accessible to the scientific community

Dynamic implementation of user-provider requirements; broad range of different types of data transactions

SIMBioMS

SAMPLE

ASSAY

AVAILABILITY Index

DATA

METADATA

Bioinformatics data standards

Data provider

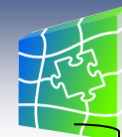
DIRECT communication on:

- project goals
- framework
- expected outcomes

Data consumer 1

Data consumer 2

Data consumer 3



# SIMBioMS

- dynamic storage
- project hosting
- fast exchange

support for collaborative discovery

stand alone researchers

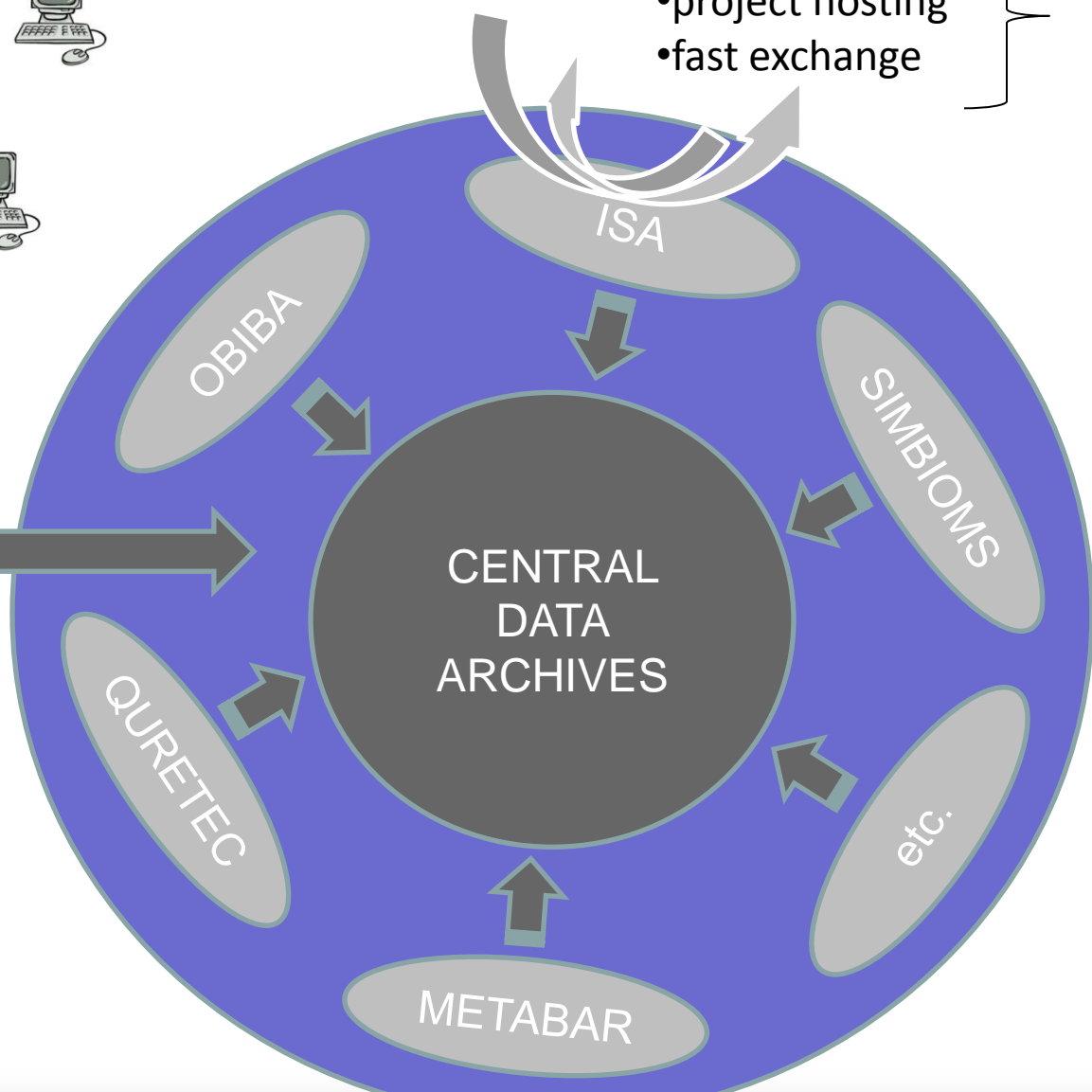


- permanent deposition
- large volumes
- open access

knowledge access and sustainability



large consortia

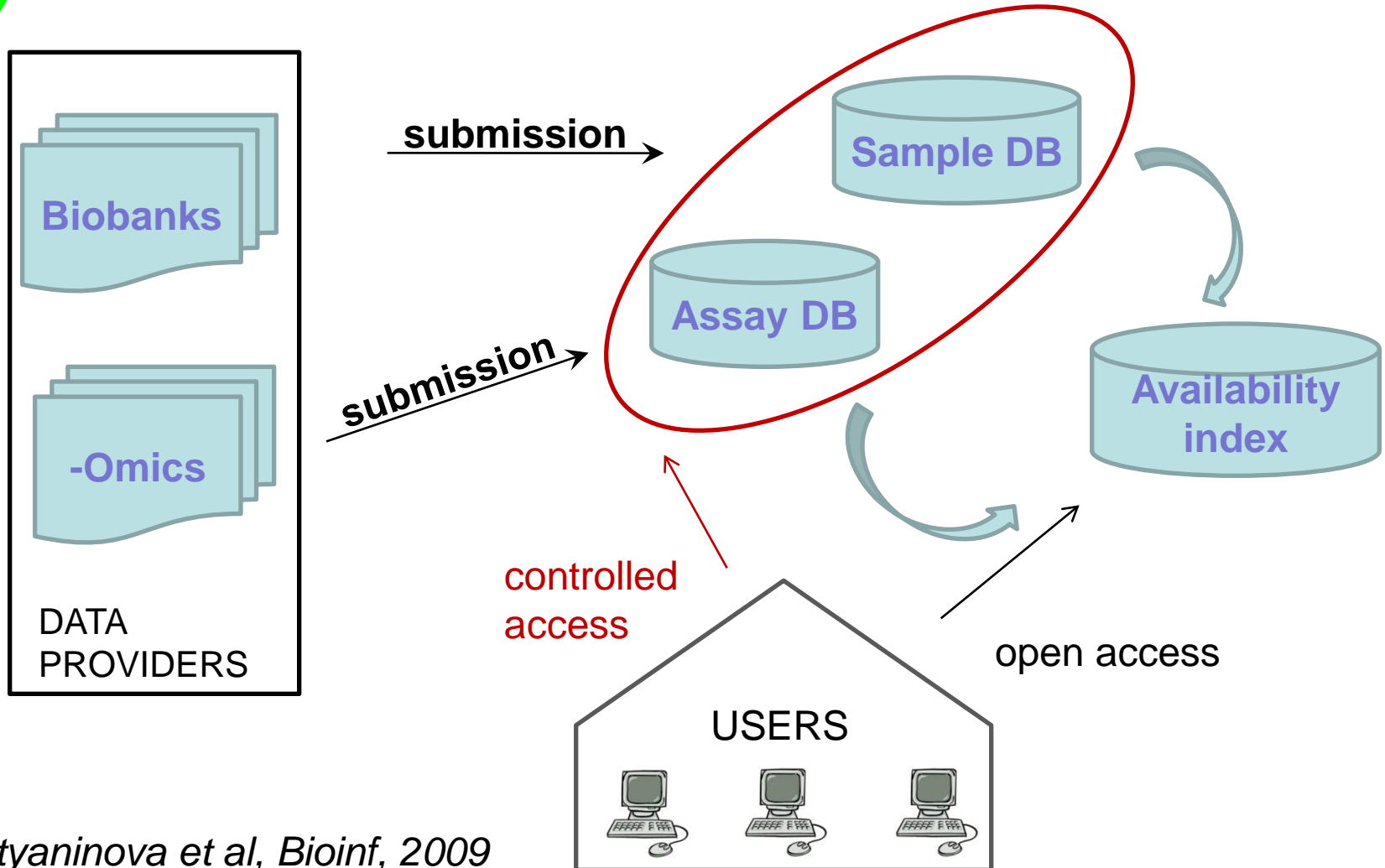


# Existing data management solutions

- [www.obiba.org](http://www.obiba.org) – open source s/w for biobanks
- [www.quiretec.com](http://www.quiretec.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*, “Optimizing data storage and exchange in collaborative research projects using adaptable open source tools” , *submitted*

# System overview



*Krestyaninova et al, Bioinf, 2009*

## 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](mailto: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, accepted*

## 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 aim is to facilitate the prospective harmonization of emerging biobanks, provide a template for retrospective synthesis and support the development of questionnaires and information 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 P<sup>3</sup>G, [CPT](#), [PHOEBE](#), and [Generation Scotland](#).



A DataSchema identifies and describes a thematic set of core variables that are of particular value in a specified scientific setting.

It also contains associated support material including variable definitions, links to relevant Ontologies and Classifications, and access to reference questionnaires and operating procedures.



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

# Current infrastructural volume

12 installation in 4 countries

>100 user-organisations

>200.000 samples

>50.000 assays and studies

4 large federated R&D across Europe  
and Russia

Krestyaninova et al, *Bioinformatics*, 2009

Viksna et al, *BMC Bioinformatics*, 2007

# Software and services

- SAIL - central index of sample and data availability
  - location, amount, context of data
  - controlled data access
- AIMS/SIMS – web-based project-specific data storage, annotation and fast exchange
  - interface with analytical tools
  - submission to public archives on request
- Emäntä - project management and coordination tracking the succession of meta-studies, e.g. from data contribution to publication
- R spider – network-based analysis of genes and molecules