

Information management in multi-partner collaborative projects.

Maria Krestyaninova, PhD
Biomedical Informatics Coordinator
EMBL-EBI, Hinxton, UK

9 February, 2010

Why informatics

- ◎ connect to peers in the same field
...and to their data...
- ◎ record keeping for posterity
...find out what you were working on a
couple of years ago
- ◎ amassing knowledge
...in case someone else knows better how
to analyse your data...

What if

you do not share global
concerns about open data
access

Practical reasons for information management

- saves time on administrative or other non-scientific tasks: project management
- reduces risks of delays in a project/study due to data logistics
- reduces risks of mishandling of data

Types of information resources

Public archives:

- www.ebi.ac.uk/ega
- <http://www.ebi.ac.uk/microarray-as/ae/>

Local systems at genetics centres:

- LIMS
- Genetic data storage

Project-specific resources

- shared ftp
- temporary database, maintained specifically for the project

What if we just want to get on with research

- ◉ project-specific communication: harmonise
- ◉ availability of samples across collections for design of a study/project
- ◉ easy transfer of large and complex datasets (genetic/molecular/medical/image)

SOLUTION I: join initiatives to enhance interconnectivity

- ◎ **Public Population Project in Genomics (P3G)**
- ◎ **Biobanking and Biomolecular Resources Research Infrastructure (BBMRI)**

ABOUT P³G

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

MEMBERSHIP

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

P³G EVENTS

- » **CHANGE OF VENUE**
P³G Annual Meeting
April 26-27, 2010
Montreal
- » P³G General Meeting
Sept 29-30, 2009
Luxembourg

WELCOME

The Public Population Project in Genomics (P³G) 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.

P³G 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

P³G 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 P³G approach.

[Click to Become a Member](#)

NEWS & P³G EVENTS

- » **CHANGE OF VENUE**
P³G Annual Meeting
April 26-27, 2010
Montreal

COMMUNICATIONS

- » [e-update](#)
January 2010
- » [Newsletter](#)
September 2009
- » [P³G e-Brochure](#)

SPONSORS





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.



SOLUTION II: find an IT platform for collaborative discovery

- ◉ highly varied level and type of in-house data management platforms
- ◉ existing public infrastructures are geared towards amassing data rather than exchange
- ◉ need for a dynamic and flexible information exchange environment to facilitate dialoguing, tracking IP ownership and to speed up the discovery

neither new, nor specific to genetics studies

...externalisation of various types of studies is expected to grow both in clinical and in pre-clinical studies.

This calls for more flexible data integration frameworks to enable various types of specialists to move data fast between sectors (pharma, CRO and biotech) as well as between knowledge domains (diagnostics, research, development)...

Questions about informatics support

- ⦿ How do we learn about the expectations?
- ⦿ How is performance evaluated?
- ⦿ Is everything to be integrated with everything?

META-LIMS solution for data intensive multi-partner interactions

- open source, web-based shared infrastructure; customisable and secure.
- designed for multi-partner projects with large and complex data
- better than ftp

- ◎ www.obiba.org – open source s/w for biobanks
- ◎ www.quiretec.com – s/w platform for data management, collection and analysis in clinical research
- ◎ <http://isatab.sourceforge.net/> – ISA (Investigation/Study/Assay) Infrastructure.

Let's look at an example

simbioms.org

In order to learn about key features of an infrastructure for fast data exchange

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

DATA

METADATA

DIRECT communication on:
• project goals
• framework
• expected outcomes

DATA
METADATA

Data consumer 1

DATA
METADATA

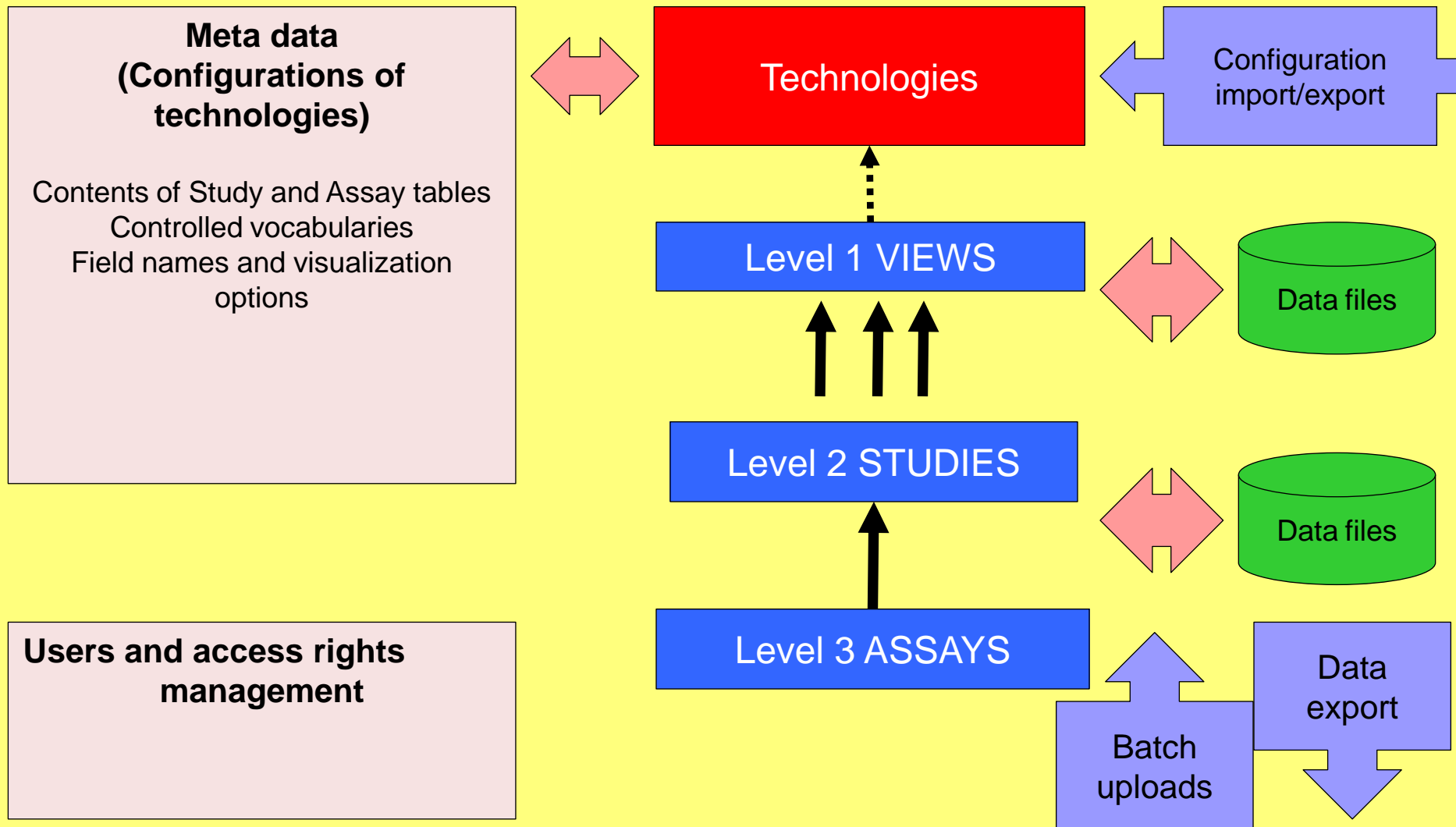
Data consumer 2

DATA
METADATA

Data consumer 3

AIMS - the basic structure

AIMS



Outline:

what does simbioms do?

- design and management of genomics studies of high complexity and high volume to facilitate collaborative discovery in the newly emerging fields

Who uses it currently?

- high scientific impact user community
- managers of high throughput data
- competitive and cooperative partnerships

how?

- Provider-to-user annotation
- ownership management
- interface with the large archives and public standards

simbioms.org

Sample Information Management System (Barcelona workshop)

List of Visits (F1)

Filter Columns User settings Batch Report Upload Visits Select Collection

1-200 of 1101 items Person

ID	FID	Gender	PC1	PC2	PC3	Type of record	Examination date	BMI	Glucose	Insulin	HOMA IR	HOMA BC	Creator	Create date			
aaa001	aaa001	2	0.00399	0.00499	-0.00378	Real data	2010-Feb-06	29.4	5.198	3.73767	-0.116	4.65871	admin	2010-Jan-20			
aaa002	aaa002	1	0.00277	-0.0124	0.01397	Real data	2010-Feb-06	28.9	5.311	3.27336	-0.56928	4.30407	admin	2010-Jan-20			
aaa003	aaa003	2	0.02699	-0.01101	0.00523	Real data	2010-Feb-06	24.4	5.198	3.70868	-0.1441	4.63861	admin	2010-Jan-20			

University of Latvia Institute of Mathematics and Computer Science

EMBL-EBI European Bioinformatics Institute

MOFA2

ENGAGE

Patient Visit Management System

2010-Jan-20

2010-Jan-20

2010-Jan-20

2010-Jan-20

Help

Add Person

List of Studies (Genotyping)

Filter Columns User settings Cancel Batch Report Data export/transfer

27 items View: Check all Uncheck all

	<u>Id</u>	Name	Date	D	Phenotype	Sample set	Gender	Direct/Imputed	Design	Subjects	Method	Chip	Genome build	Type	Ad
<input checked="" type="checkbox"/>	ADM-GT-1	Test1	2009-Dec-29		cholesterol	NTR	F	direct	whole sample set		Perlegen	Affymetrix Genome-...	35	standard	z-s adj
<input checked="" type="checkbox"/>	ADM-GT-2	F1	2010-Jan-12		Fasting Glucose	F1	M+F	direct	whole sample set	1101	Illumina	Illumina HumanHap370	36	standard	nor
<input checked="" type="checkbox"/>	ADM-GT-3	F2	2010-Jan-12		Fasting Glucose	F2	M+F	direct	whole sample set	1101	Illumina	Illumina HumanHap370	36	standard	nor
<input type="checkbox"/>	ADM-GT-4	F3	2010-Jan-12		Fasting Glucose	F3	M+F	direct	whole sample set	1101	Illumina	Illumina HumanHap370	36	standard	nor
<input type="checkbox"/>	ADM-GT-5	F4	2010-Jan-12		Fasting Glucose	F4	M+F	direct	whole sample set	1000	Illumina	Illumina HumanHap370	36	standard	nor
<input type="checkbox"/>	ADM-GT-6	F5	2010-Jan-12		Fasting Glucose	F5	M+F	direct	whole sample set	1100	Illumina	Illumina HumanHap370	36	standard	nor
<input checked="" type="checkbox"/>	ADM-GT-7	F1	2010-Jan-12		Fasting Glucose	F1	M+F	direct+imputed(HM2)	whole sample set	1101	Illumina	Illumina HumanHap370	36	standard	nor
<input checked="" type="checkbox"/>	ADM-GT-8	F2	2010-Jan-12		Fasting Glucose	F2	M+F	direct+imputed(HM2)	whole sample set	1101	Illumina	Illumina HumanHap370	36	standard	nor
<input type="checkbox"/>	ADM-GT-9	F3	2010-Jan-12		Fasting Glucose	F3	M+F	direct+imputed(HM2)	whole sample set	1101	Illumina	Illumina HumanHap370	36	standard	nor
<input type="checkbox"/>	ADM-GT-10	F4	2010-Jan-12		Fasting Glucose	F4	M+F	direct+imputed(HM2)	whole sample set	1000	Illumina	Illumina HumanHap370	36	standard	nor
<input type="checkbox"/>	ADM-GT-11	F5	2010-Jan-12		Fasting Glucose	F5	M+F	direct+imputed(HM2)	whole sample set	1100	Illumina	Illumina HumanHap370	36	standard	nor

Provision of data management services to R&D partnerships:

- fast exchange +long-term deposition (tutorial)
- ownership rights (lecture)
- rich annotation of data files (
- simple deployment and maintenance (lecture)

Current infrastructural volume

- 12 installation in 3 countries
- 100 user-organisations
- >50.000 samples
- >50.000 assays and studies
- 4 large federated R&D projects across Europe and Russia

What if we just want to get on with research

- ◉ project-specific communication: harmonise
- ◉ availability of samples across collections for design of a study/project
- ◉ easy transfer of large and complex datasets (genetic/molecular/medical/image)

SAIL: Sample avAILability

- System for browsing and annotating large and complex data across many sources
- Combine samples based on experimental parameters and variables measured
- Sharing availability information for phenotypes across cohorts for meta-analysis study planning



http://www.ebi.ac.uk/Tools/sail/

Collection filter

Request field

Welcome Summary Report constructor

Study: [ANY] Collection: [ANY]

Parameter list Parameter tree Parameter hierarchy

Vocabulary MetS Columns

Code	Name	Description	Filter	Re...	V	E
AGE	Age	Age		33080	1	0
ALC				13741	1	1
ALCQ	Alcohol quantity	grams absolute ethanol / ...		16252	1	0
	Antihypertensives	Antihypertensive treatment		11677	1	1
APOB	Apo B mg/l	Biochemistry Apolipoprotein B		1781	1	1
BMI	BMI	Body Mass Index, kg/m ²		32569	1	0
BASO	BASO	Blood Basophils		69	1	0
BICEPS	Biceps mm	Thickness of a skinfold on...		69	1	0
BYR	Birth Year	Birth Year		14994	1	0
BP	Blood pressure	Blood pressure (systolic)		22412	2	0

Tag filter

Text search

Value filtering tool

Study filter

Phenotypes

Report request

AND OR

Request

Use split by collection

Use re

Specify collections

Specif

Select

Quick query Add Add to group Relations Extra

Query

Availability report

Summary

Report constructor

Report 2

Total records: 33102

Number of samples that have both glucose and insulin measurements for "UK Twins" cohort

Collections	Records in collection	GLU ¹	INS ²	
UK Twins	6008	5372	4360	4279
ERF	3205	3205	2256	2256
Summary	9213	8577	6616	6535

Potential sample size

¹ - parameter with code: 'GLU' and name: 'Glucose'

² - parameter with code: 'INS' and name: 'Insulin'

³ - GLU **AND** INS

simbioms.org

Cohort summary

KoraF4

Summary info

Total samples: 1814
 Samples with tag 'Genotyped': 1814
 Last update: 23 April 2009 00:00:00

Country

Germany

Owner

German Research Center for Environmental Health (HZM, former GSF)

Contacts

Christian Gieger christian.gieger@helmholtz-muenchen.de
 Melanie Kolz melanie.kolz@helmholtz-muenchen.de
 Tech.contact: Janina Ried janina.ried@helmholtz-muenchen.de

Measured parameters

Parameter	Variable or Qualifier		Records	Genotyped
	Variable or Qualifier	Variant		
Gen:GENDT (Genotyped)			1814	1814
RESID (Residence)	Country	Germany	1814	1814
SEX (Sex)	Sex	Man	884	884
		Woman	930	930
AGE (Age)			1814	1814
EXYR (Year examination)			1814	1814
EDU (Education)			1811	1811
SMK (Smoking status)	Status	no	1582	1582
		yes	230	230
SMKQ1 (Smoking quantity 1)			1812	1812

http://www.ebi.ac.uk/Tools/sail/SAIL_USER.html

Content of SAIL*

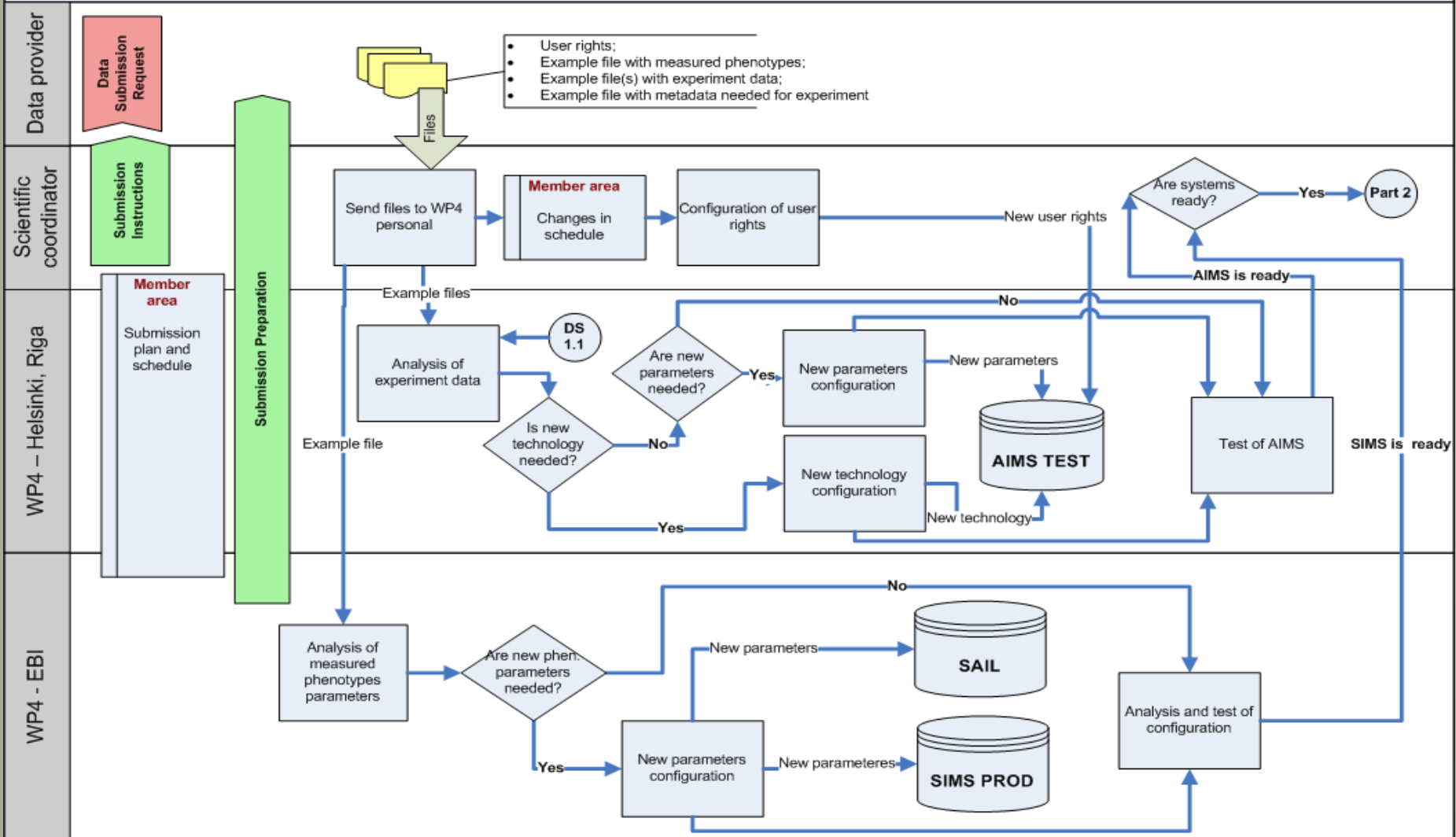
- ① >30,000 samples
- ① 10 cohorts (biobanks)
- ① 83 variables harmonised across 10 biobanks;
51 harmonised across 6 twin collections

(*) Manuscript in preparation, available on request

Cross-cohort availability index in meta studies

- In ENGAGE, meta-analysis studies are designed to increase power to detect genotype-phenotype associations
- SAIL enables data owners to design studies and maximize sample sizes without having to share individual data

Data Submission Process in ENGAGE, part 1



Conclusions

- From centralised data amassing to a network of project- specific exchange hubs
- Informatics is a part of cost-recovery model
- From data volume to data demand oriented architectures:
 - N of data transactions ~ knowledge gain