Session 2: Data Archiving and Management: the IT Infrastructure
Chair: Jan-Eric Litton, Karolinska Institutet, SE

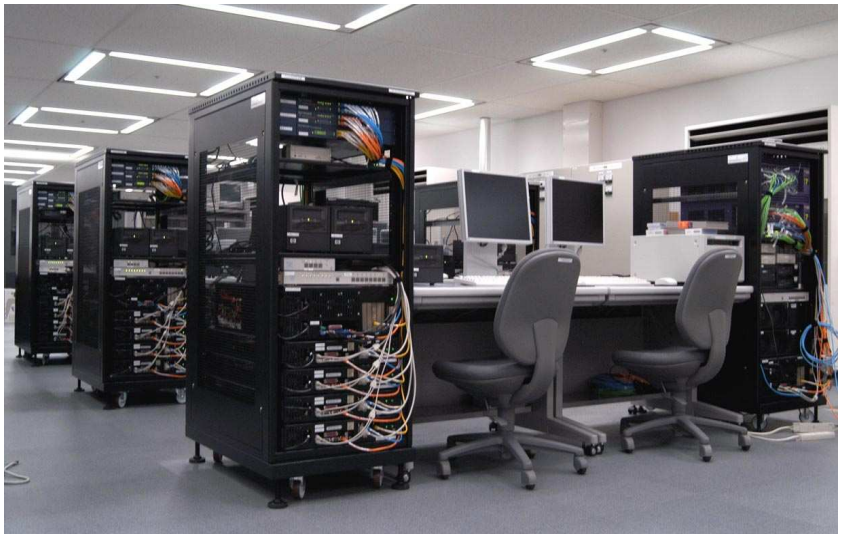
15.00-15.40 Erich Wichmann
Helmholtz Center, DE
Phenotypic and genetic information associated to a biobank
15.40-16.20 Juha Kononen
University of Tampere, FI
Creation of integrated data mining environment linking individual data to clinical cellular and molecular information

16.20-16.50 Coffee break
16.50-17.30 J.J. Nietfeld
University Medical Center Utrecht and INTRESCO, NL
Software codes versus biocodes for sample and data traceability
17.30-18.10 Jan-Eric Litton
Karolinska Institutet, SE
Harmonization of IT infrastructure across countries
18.10-18.25 Inés Barrecheguren
Noray Bioinformatics, S.L., ES
Software platforms for Biobanks Information management and traceability
18.25-18.40 Kharlampi Tiras
Russian Academy of Sciences, RU
General biology in 21 century: from museums of "mortified" samples to electron images of living biological objects
19.00 Dinner

Biobank

Not only biological samples…

.. but also an information management system



Harmonization of IT infrastructure across countries

Prof. Jan-Eric Litton
Karolinska Institutet
Stockholm, Sweden.

jan-eric.litton@ki.se

Agenda

- The problems to solve …
- Requirements
- Variable explosion
- The Federated database

- GenomEUtwin and TwinNET
- Bio-GRID
- BIMS

- BBMRI

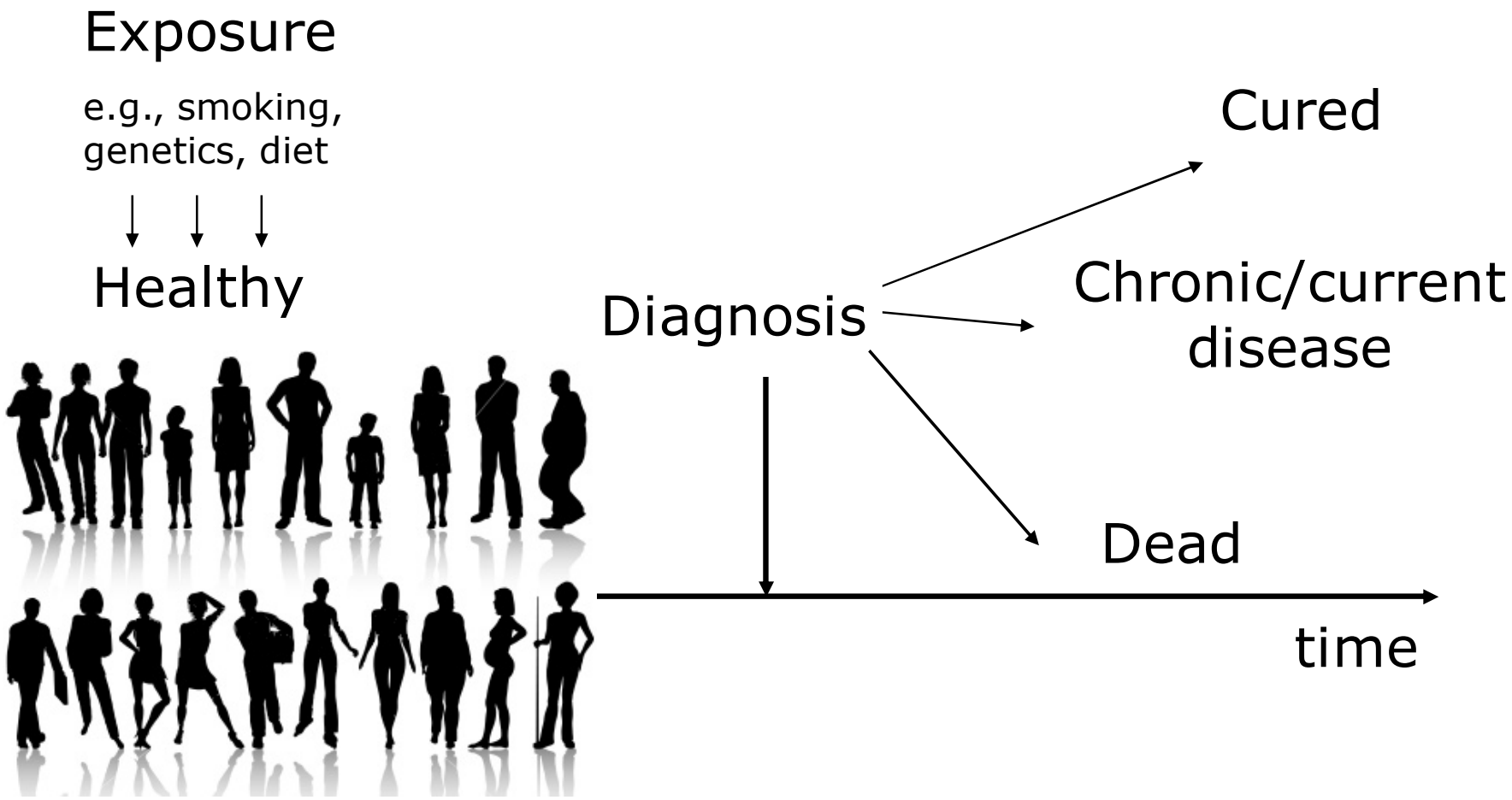
World Wide Biobanking

General Requirements:

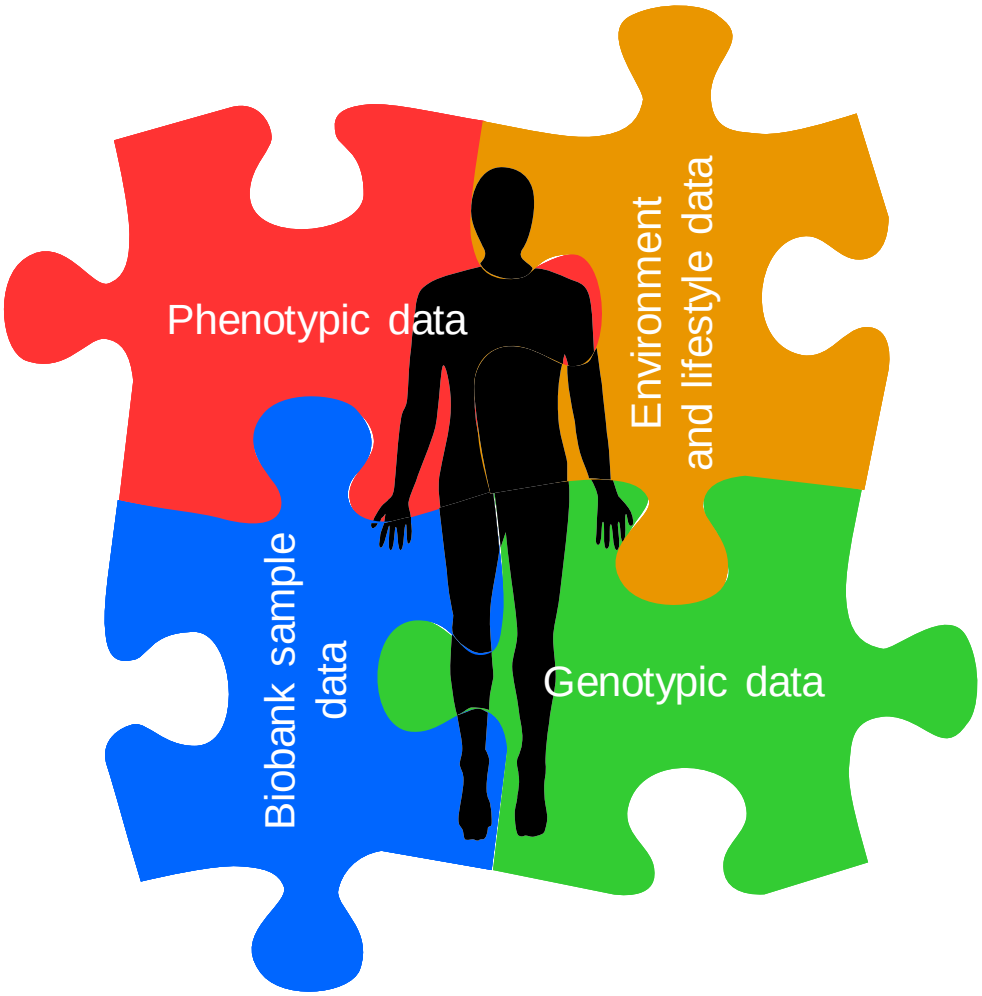
1. ID
2. Data model - Meta data
3. Nomenclature

Simplexity = Simple Rules of a Complex World

Popultaion Biobanks



Requirements

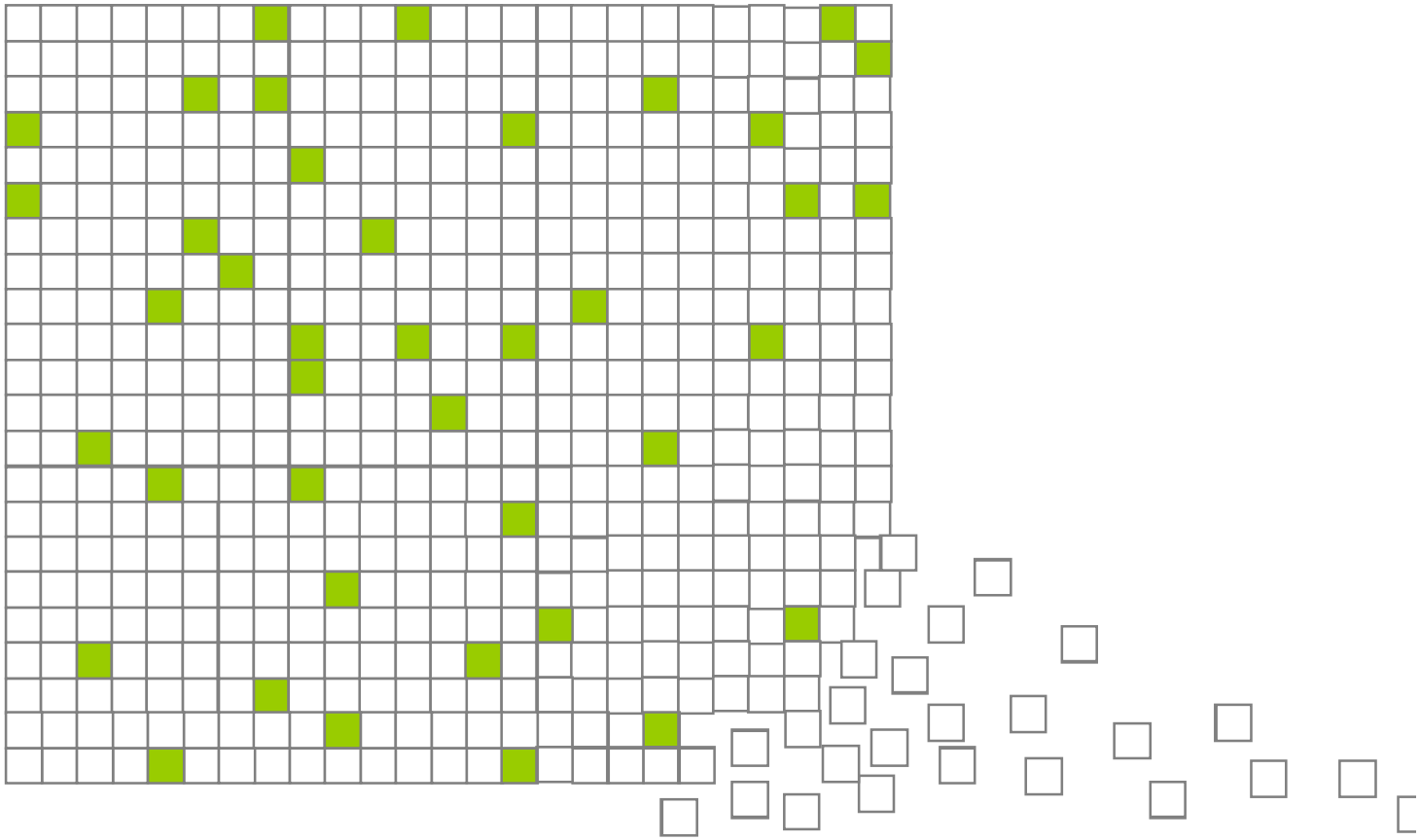


IT- Requirements:

1. Store Large Amounts of Genotype Data

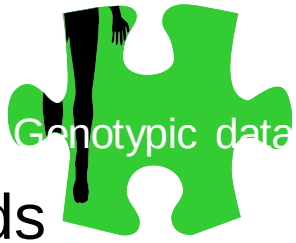
n_{SNP} = 1 000 000

n_{person} = 1 000 000



IT- Requirements:

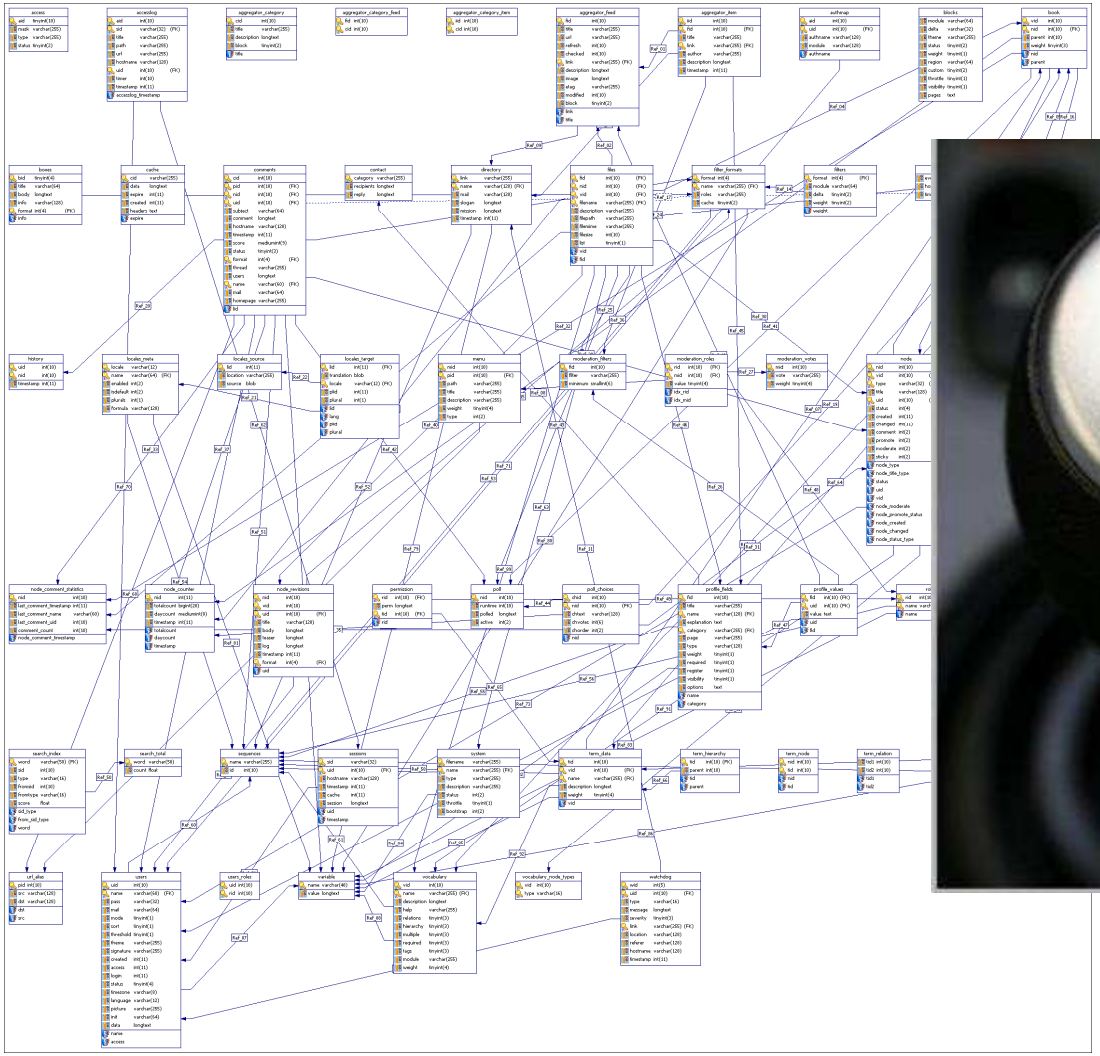
2. Fast Extraction of Genotype Data



IT- Requirements:

3. Handle Datasets Analyzed with Different Methods

IT- Requirements:
4. Include Meta-Data



Genotypic data

IT - Requirements:
5. Scalable and Cost-Effective

Factors driving variable explosion

- Methods for high-throughput single nucleotide polymorphism (SNP) genotyping analysis continue to improve (and becoming less expensive)
  - Illumina human1M beadchip
  - Affymetrix® Genome-Wide Human SNP Array 6.0

- Gene-environment interaction studies in epidemiology are constantly increasing in size
  - LifeGene
  - UK Biobank
  - Singapore

- The huge amount of data generated when high-throughput SNP genotyping methods are used in such studies, presents an enormous challenge to researchers in terms of structured data management

Agenda

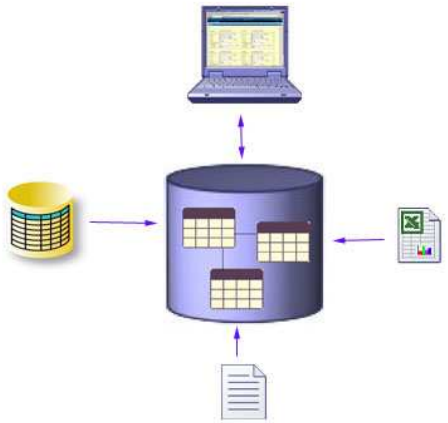
- The problems to solve …
- Requirements
- Variable explosion
- The Federated database

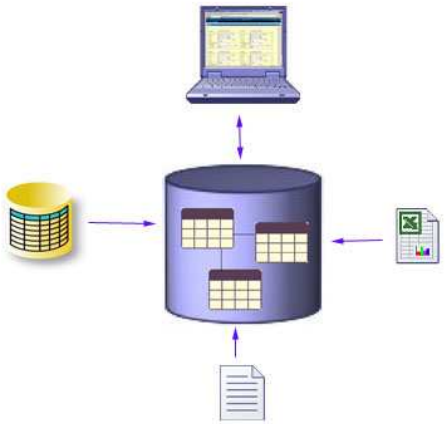
- GenomEUtwin and TwinNET
- Bio-GRID
- BIMS

- BBMRI

A federated database system is a type of meta-database management system (DBMS) which transparently integrates multiple autonomous database systems into a single federated database.

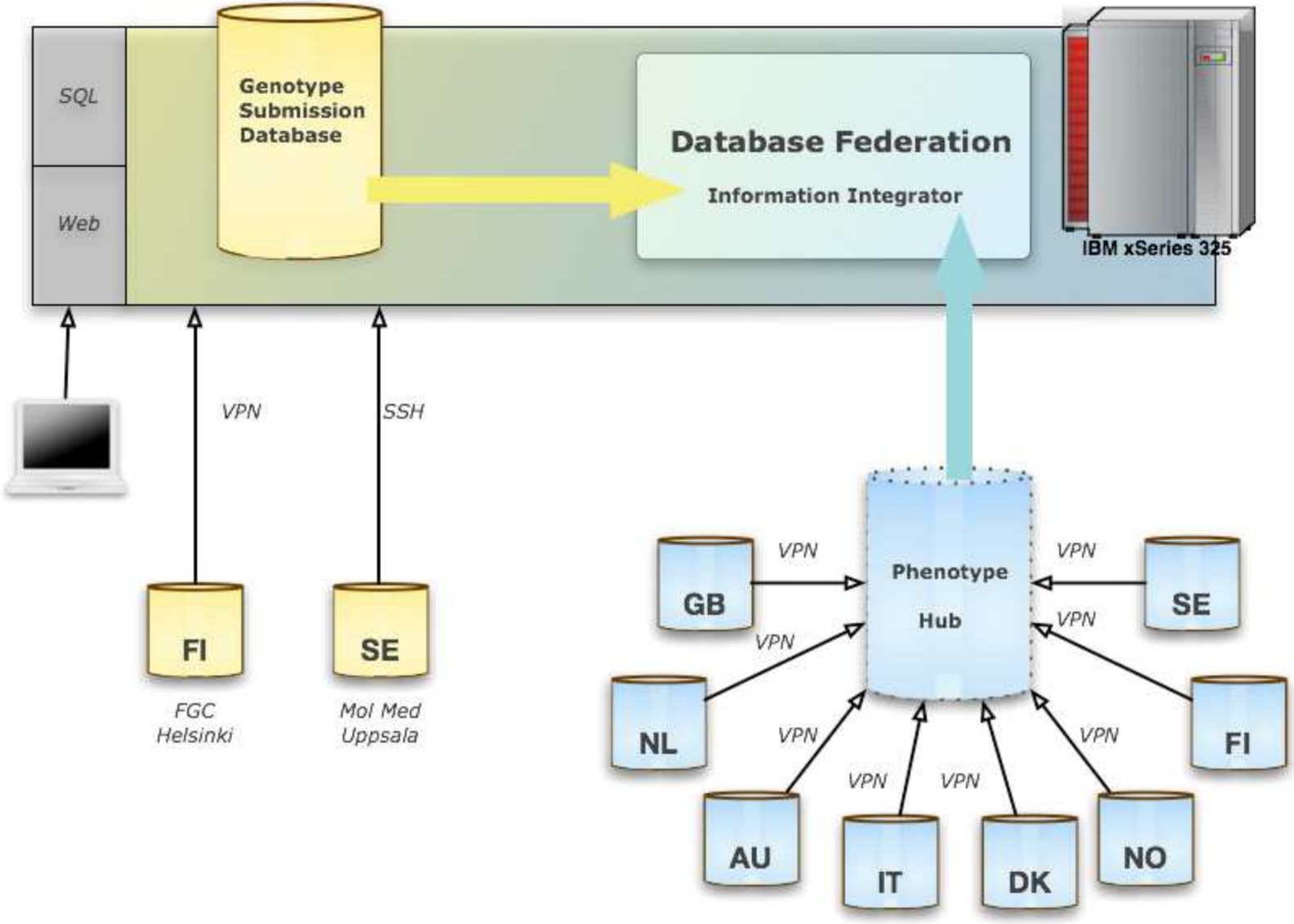
The constituent databases are interconnected via a computer network, and may be geographically decentralized.





A federated database (or **virtual database**) is the fully-integrated, logical composite of all constituent databases in a federated database system.

A DBMS can be classified as either centralized or distributed. A centralized system manages a single database while distributed manages multiple databases.

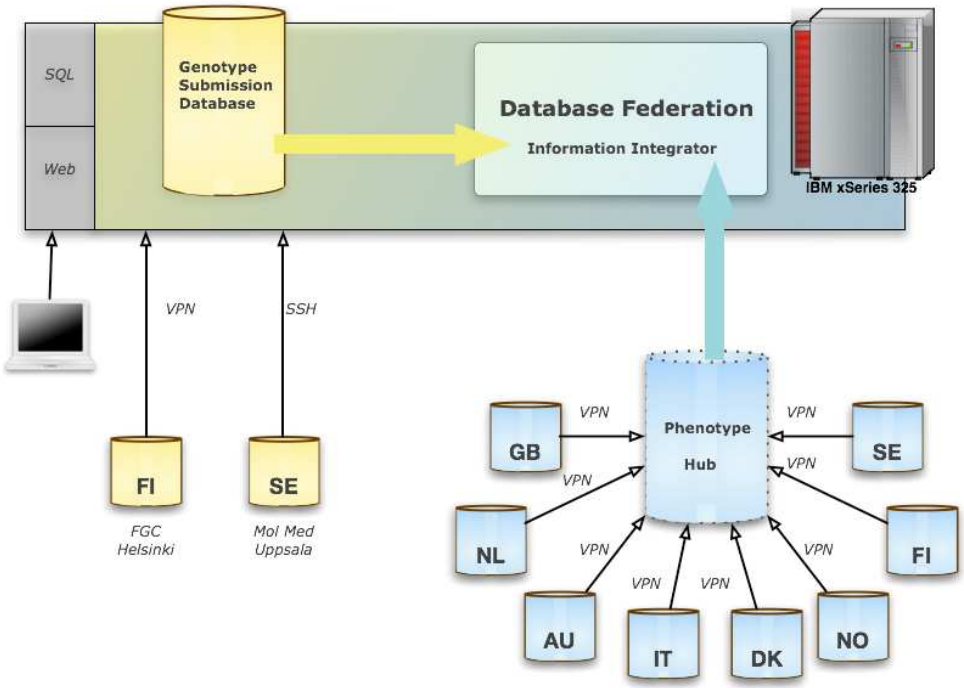


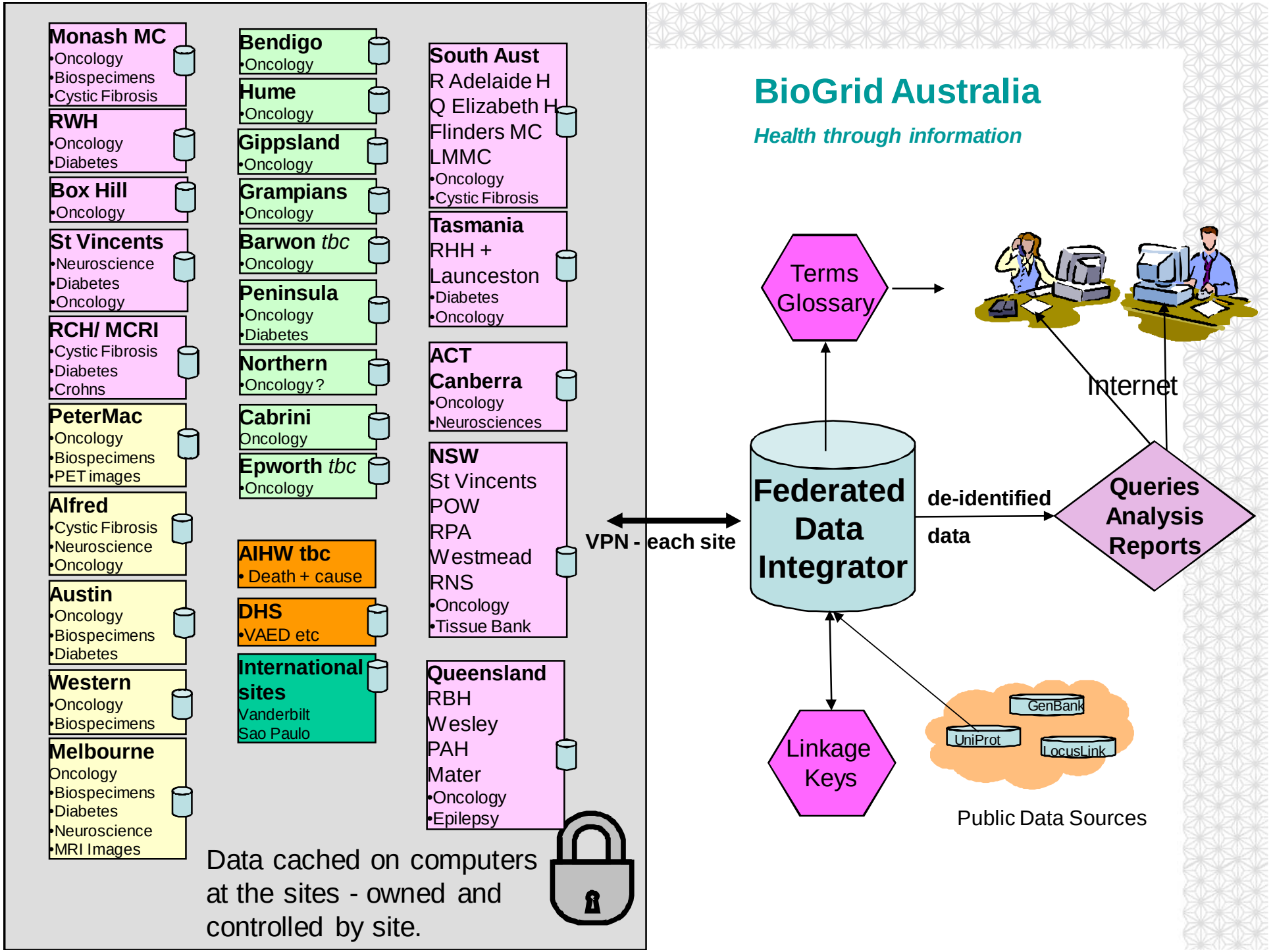
Muilu J, Peltonen L**, Litton JE**. The federated database - a basis for biobank-based post-genome studies, integrating phenome and genome data from 600 000 twin pairs in Europe. Eur J Hum Genet, 2007, **15,** 718Š723

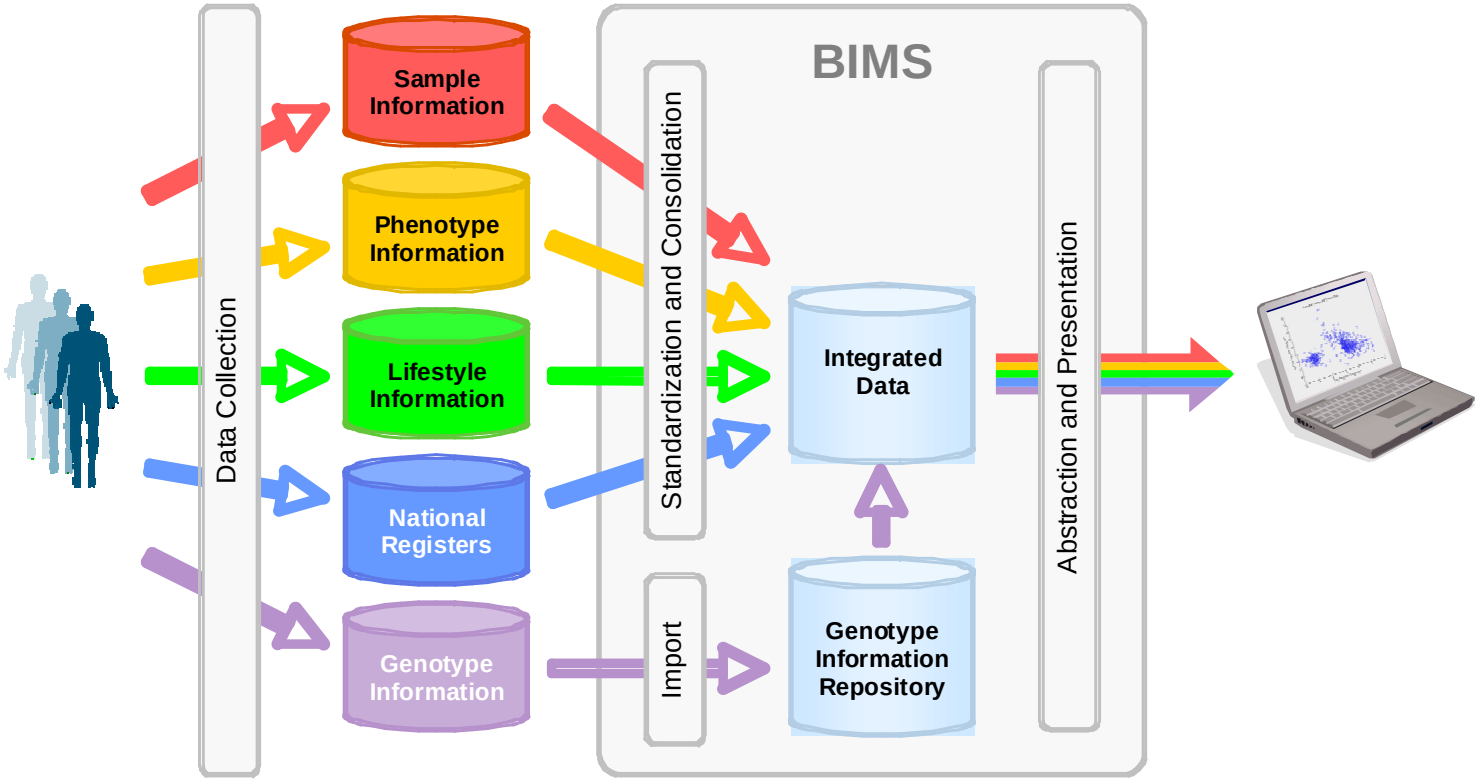
Data Format and Variable Standard for GenomEUtwin's Phenotype Database

Version: 3.2.4

Ann Björklund
Jan-Eric Litton




Monash MC
•Oncology
•Biospecimens
•Cystic Fibrosis
RWH
•Oncology
•Diabetes
Box Hill
•Oncology
St Vincents
•Neuroscience
•Diabetes
•Oncology
RCH/ MCRI
•Cystic Fibrosis
•Diabetes
•Crohns
PeterMac
•Oncology
•Biospecimens
•PET images
Alfred
•Cystic Fibrosis
•Neuroscience
•Oncology
Austin
•Oncology
•Biospecimens
•Diabetes
Western
•Oncology
•Biospecimens
Melbourne
Oncology
•Biospecimens
•Diabetes
•Neuroscience
•MRI Images
Bendigo
•Oncology
Hume
•Oncology
Gippsland
•Oncology
Grampians
•Oncology
Barwon tbc
•Oncology
Peninsula
•Oncology
•Diabetes
Northern
•Oncology ?
Cabrini
Oncology
Epworth tbc
•Oncology
AIHW tbc
• Death + cause
DHS
•VAED etc
International sites
Vanderbilt
Sao Paulo
South Aust
R Adelaide H
Q Elizabeth H
Flinders MC
LMMC
•Oncology
•Cystic Fibrosis
Tasmania
RHH +
Launceston
•Diabetes
•Oncology
ACT
Canberra
•Oncology
•Neurosciences
NSW
St Vincents
POW
RPA
Westmead
RNS
•Oncology
•Tissue Bank
Queensland
RBH
Wesley
PAH
Mater
•Oncology
•Epilepsy
Data cached on computers at the sites - owned and controlled by site.
VPN - each site
Federated Data Integrator
de-identified data
Queries Analysis Reports
Terms Glossary
Internet
Linkage Keys
GenBank
UniProt
LocusLink
Public Data Sources
BioGrid Australia
Health through information



..lund G, Lindqvist P**, Litton J-E**. BIMS: An information management system for biobanking in the 21st century. IBM Systems Journal 2007;46(1):171-182.

Agenda

- The problems to solve …
- Requirements
- Variable explosion
- The Federated database

- GenomEUtwin and TwinNET
- Bio-GRID
- BIMS

- BBMRI

BBMRI
Biobanking and Biomolecular Resources Research Infrastructure

>200

Construction of new infrastructures - preparatory phase

FP7-INFRASTRUCTURES-2007-1

BBMRI
Biobanking and Biomolecular Resources Research Infrastructure

Ideally, BBMRI should build something that every one needs, but no-one has.

BBMRI
Biobanking and Biomolecular Resources Research Infrastructure

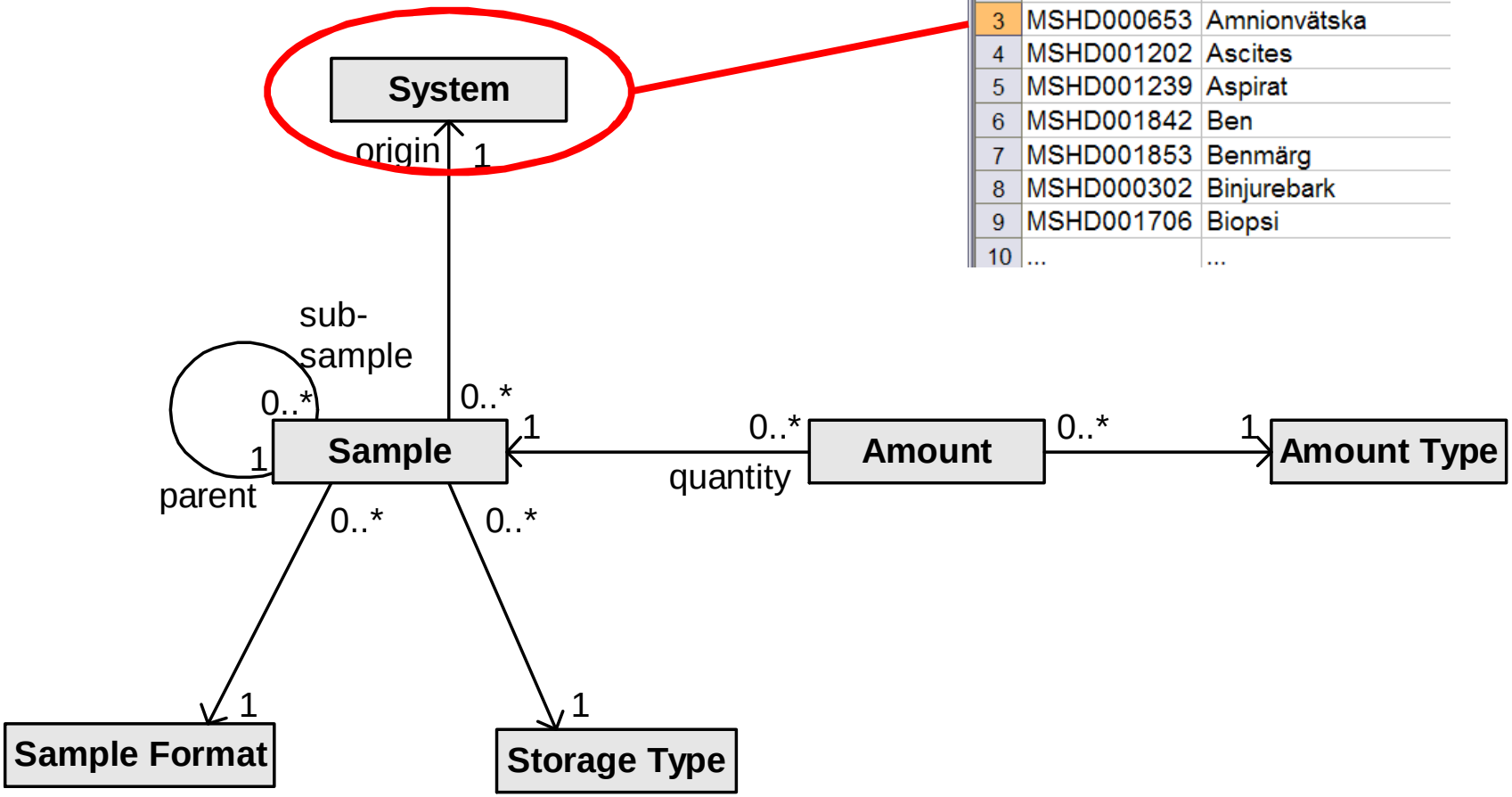
Ideally, BBMRI should build something that every one needs, but no-one has.

Searching the biobank domain in Europe

BBMRI
Biobanking and Biomolecular Resources Research Infrastructure

Biobank entities -physical

	Kod	System
2		Abcess
3	MSHD000653	Amnionvätska
4	MSHD001202	Ascites
5	MSHD001239	Aspirat
6	MSHD001842	Ben
7	MSHD001853	Benmärg
8	MSHD000302	Binjurebark
9	MSHD001706	Biopsi
10	...	...



BBMRI
Biobanking and Biomolecular Resources Research Infrastructure

META - data

A metadata system, with ranking of results, for biobanks should be defined for three levels:

- Semantic structure and metadata exploration
- Aggregating/OLAP (Online analytical processing) level (i.e., existence of data)
- Object level (e.g., samples, individuals)

A minimum set of information must be defined for each metadata level

Nomenclature - Wiki

Biobanks Wiki

biobank-lexicon.org and biobank-lexicon.com

OWL/OBO formats

English, Swedish, Estionian and Finish definitions

Use Cases for Federated Biobanks

Content

- Workflow of Use Cases for Federated Biobanks

- Use Case 1.

Identification of Biobanks

Example request for Use Case 1b

*A researcher wants to know which biobanks store **at least 500 paraffin tissues** of **mamma carcinoma** together with **therapy description**.*

- Inputparameter:
  - Paraffin
  - Tissue
  - Mamma carcinoma e.g C50.08
  - Therapy description available

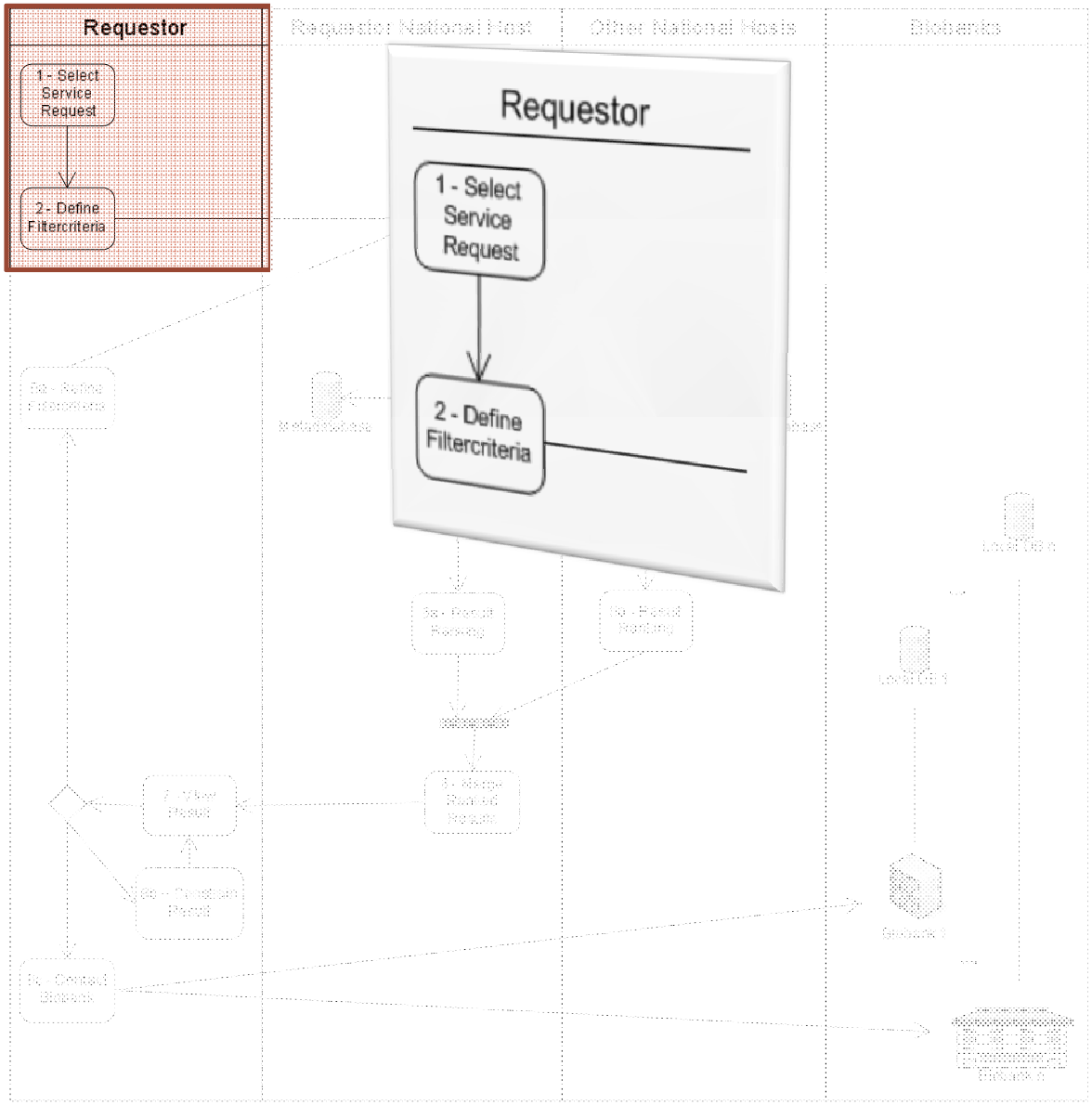
Approximate available amount indicated as:
- Few (> 2 & <= 100)
- ***Some (> 100 & <= 500)***
- Many (> 500 & <= 1000)
- Very many(> 1000)

- Meta-Database has knowledge about approximate amount of items for defined sets of attributes

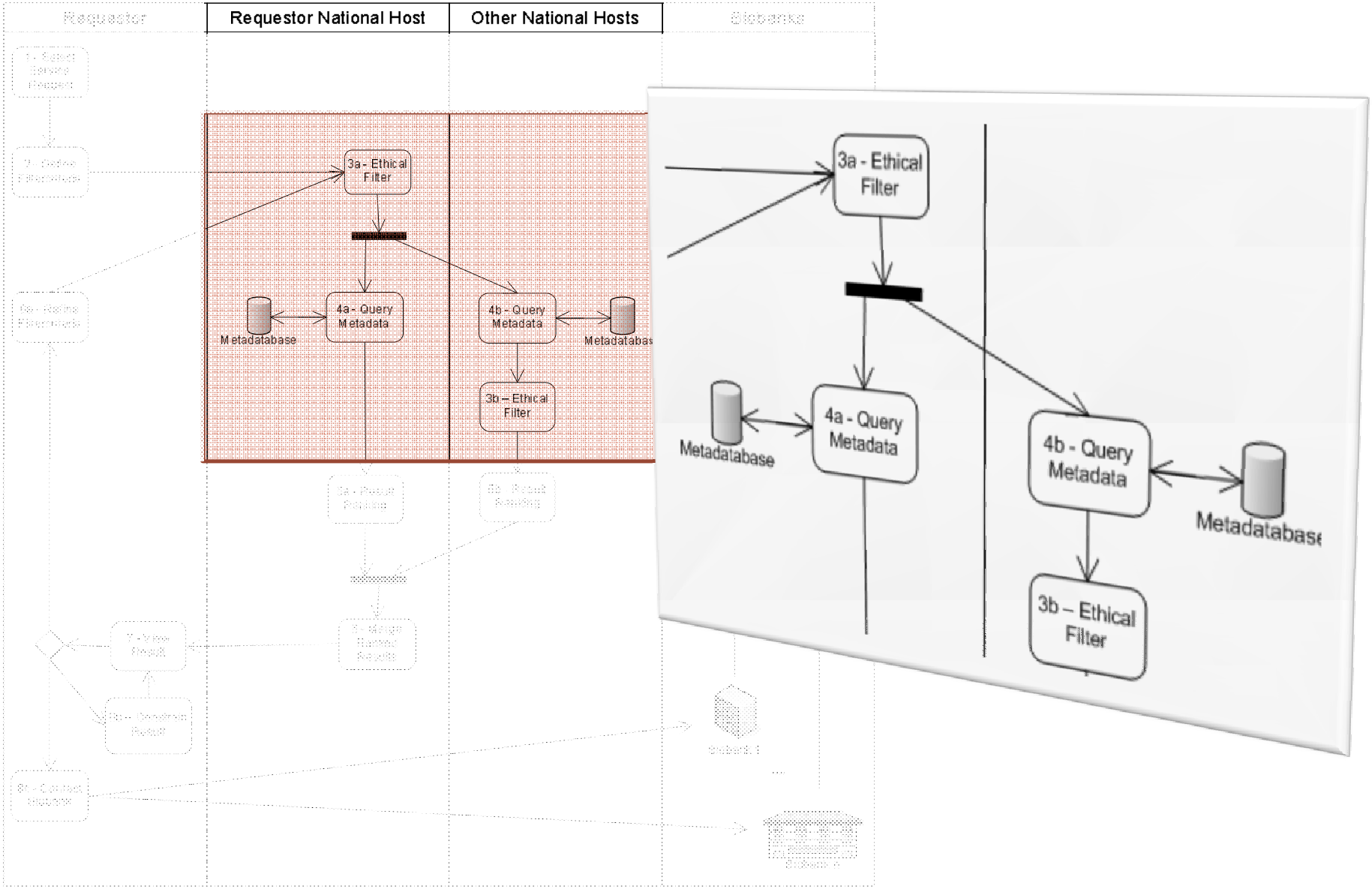
Data in Meta-Database

Biobank	Disease	Sample	Preparation / Quality	Order of magnitude
bb1	C50.08	Tissue	Paraffin	250
bb2	C50.08	Blood	Cryo	700
bb3	C50.08	Tissue	Paraffin	400
…	…	…	…	…

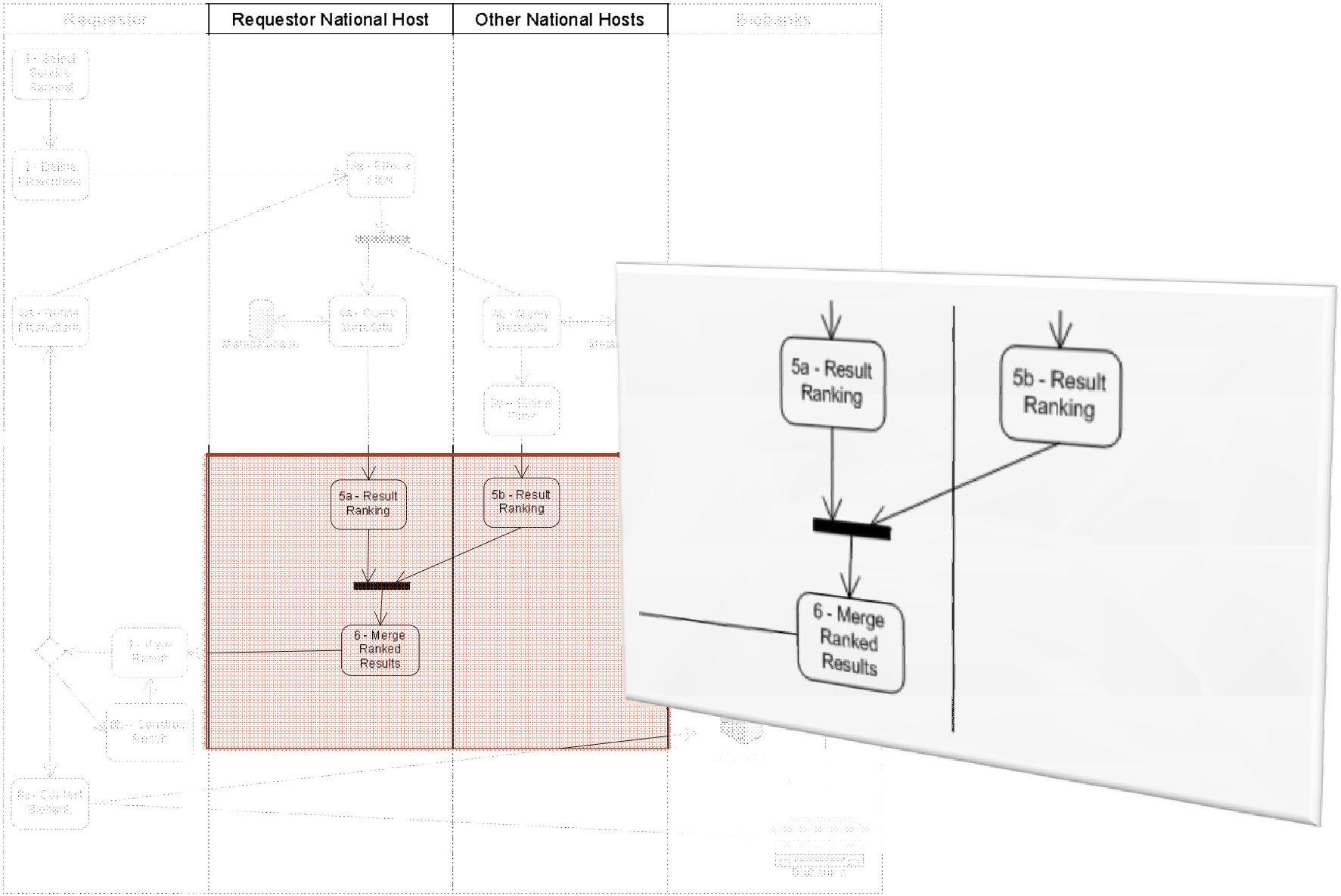
Use Case 1: Identification of Biobanks (i)



Use Case 1: Identification of Biobanks (ii)



Use Case 1: Identification of Biobanks (iii)

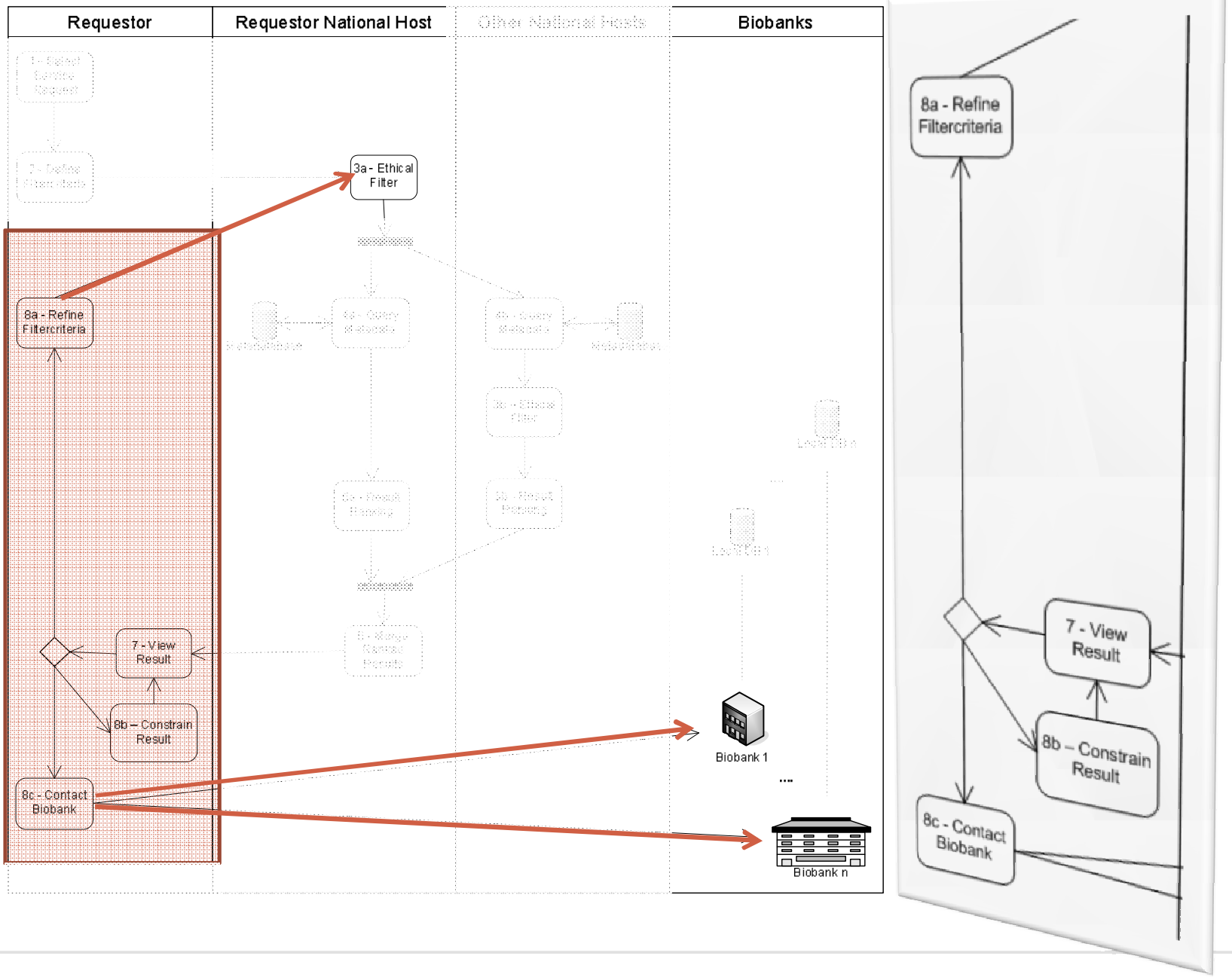


Step 5 – Result Ranking (i)

- Result Ranking is split up into step 5a and 5b which work in different environments resp. national hosts.
- Resulting data from *query metadata* is sorted according to predefined importance of attributes
- The more likely the result and the importance template are, the more important the result is treated

- **Output description:** List of *ranked* biobanks (associated with national hosts) which hold the requested data or material based on the specified attributes. Rank can be any number between 0 and 100.

Use Case 1: Identification of Biobanks



Step 7 – View Result

- The ranked list of biobanks is shown to the user
  - Available attributes / items for each biobank are shown
- Requestor can choose between three ways to proceed:
  - Step 8a - Refine filtercriteria: Query parameters can be adapted and query is sent once again.
  - Step 8b - Constrain result: Requestor can constrain the result on attributes that are in the resultset.
  - Step 8c - Contact biobank: Requestor can use biobank contact information to get in touch with a biobank.

jan-eric.litton@ki.se