GEN-AU
GENOMFORSCHUNG IN ÖSTERREICH

Data Management for Biobanks

JOHANN EDER

CLAUS DABRINGER
MICHAELA SCHICHO
KONRAD STARK

University of Klagenfurt and University of Vienna

Ein Programm des BMWF Programmmanagement: FFG Wissenschaftskommunikation: science communications

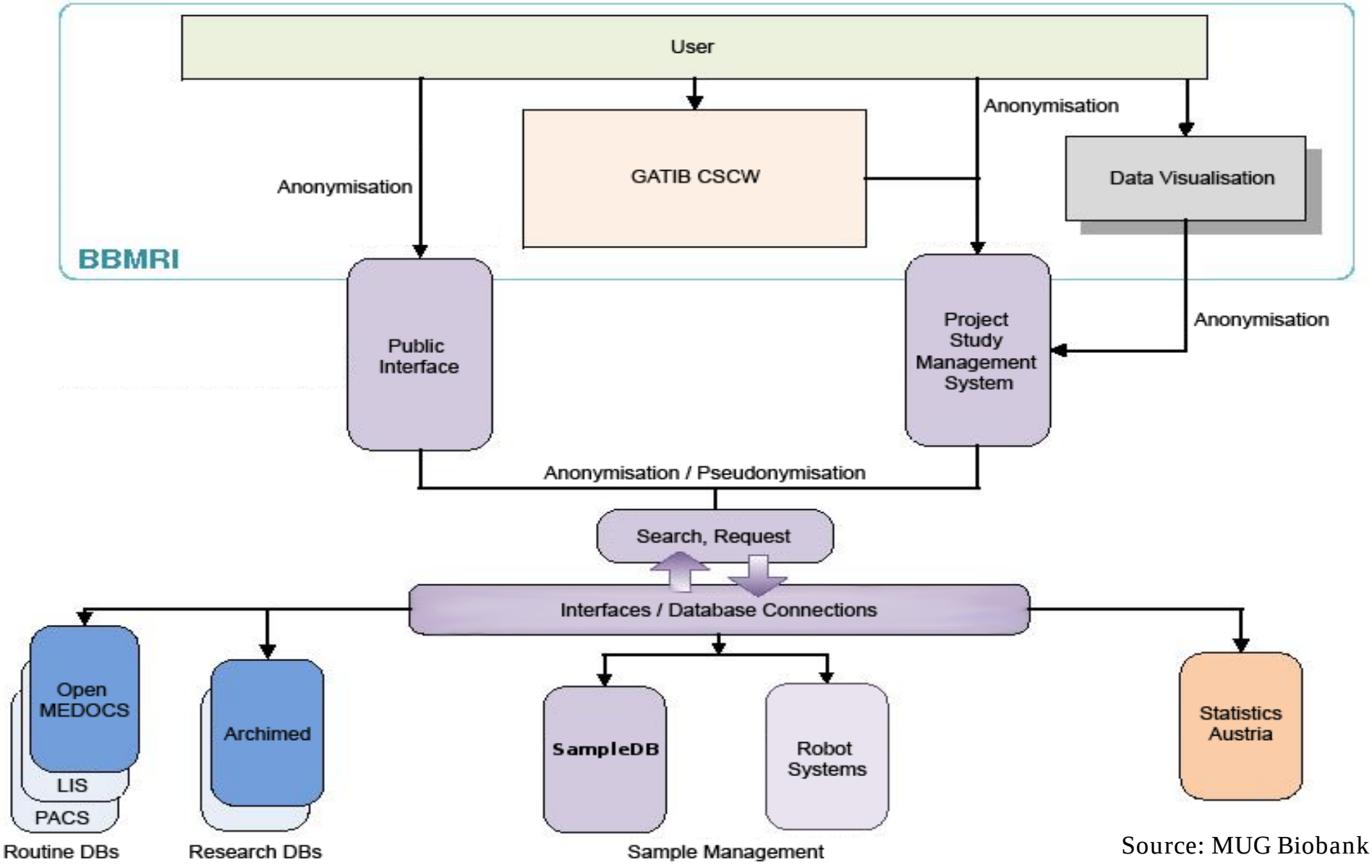
GEN-AU
GENOMFORSCHUNG IN ÖSTERREICH

Data Management for Biobanks

- Local Integration
- Project Support
- Anonymization of Sensitive Data
- Global Integration

Ein Programm des BMWF Programmmanagement: FFG Wissenschaftskommunikation: science communications

Local Integration



Source: MUG Biobank

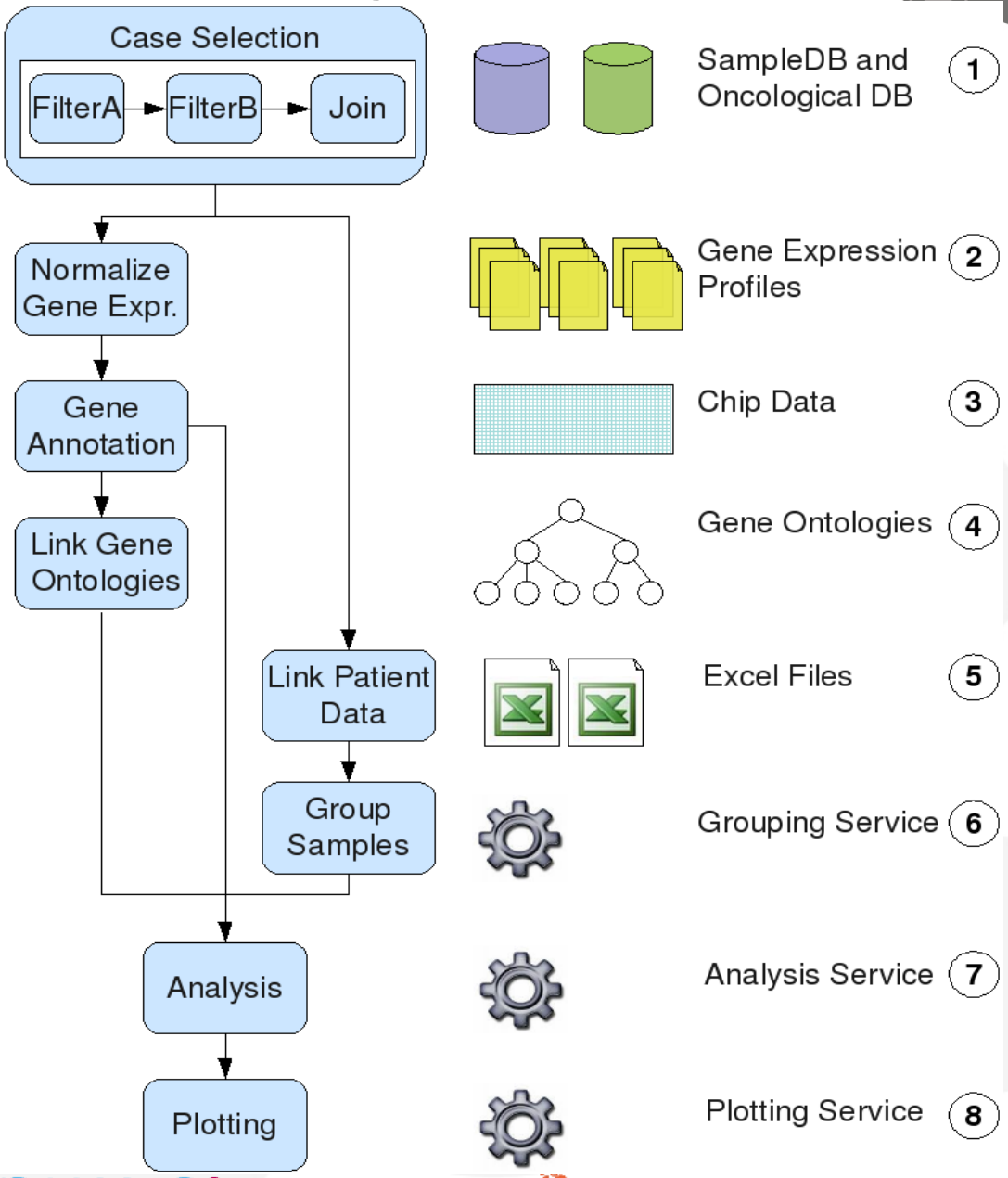
Project Support for Medical Research

- Computer Supported Collaborative System
  - Medical research is a highly complex collaborative process
  - Interdisciplinary environment (medical, biological and technical scientists)
  - Distributed resources: patient records, images, gene expression profiles, survival data, lifestyle data
  - Different access rights and permissions (Privacy of patient records, k-anonymization for external cooperation)
  - Cooperative data annotation, formulate hypotheses, statistical analyses
  - Development of a *Virtual Scientific Environment* allowing to share data and knowledge among collaborating groups

Main Requirements of a CSCW System

- R(1) User and Role Management
- R(2) Transparency of physical storage
- R(3) Flexible data presentation
- R(4) Flexible integration and composition of services
- R(5) Support of cooperative functions
- R(6) Data-coupled communication mechanisms
- R(7) Knowledge creation and transparency

Gene Expression Analysis Workflow



Anonymization of sensitive-data

- Release of person-specific data
- Protection of the anonymity of the individuals
- Data anonymization
  - Eliminate explicit identifying attributes
  - Transformation of attribute values (generalisation, recoding,..)
- Problems
  - Information loss: Transformed data may not be applicable in further processing
  - Data quality: Different data quality decrease in transformations

Anonymisation of patient data – Our approach

- Released records meet the k-anonymity constraint
- Anonymization is influenced by data quality requirements
- Search for appropriate anonymization solution:
  - Weighted generalisation hierarchies
  - User-defined priorities and generalisation limits
- Achieve k-anonymity in distributed sources
  - Generate data twins for each data source separately
  - Consider selection and projection criteria

Example: Data to be released

Staging T	Staging N	Staging M		Cause of Death	Day of Death		Height(cm)	Weight(kg)
3A	0	1		Atherosclerotic heart disease	08.03.2003		185	85
3A	0	X		Sigmoid colon	07.12.1999		173	80
3B	0	X		Colon, unspecified	22.06.2000		175	69
3B	0	X		Sigmoid colon	01.09.1997		172	70
3B	0	1		Aterosclerotic heart disease	06.03.1999		170	62
3B	0	X		Caecum	12.11.2001		183	80

Example: Transformations

GEN-AU
GENOMFORSCHUNG IN ÖSTERREICH

Generation of data twins

10

Staging T	Staging N	Staging M		Cause of Death	Day of Death		Height(cm)	Weight(kg)
3A	0	1		Atherosclerotic heart disease	08.03.2003		185	85
3A	0	X		Sigmoid colon	07.12.1999		173	80
3B	0	X		Colon, unspecified	22.06.2000		175	69
3B	0	X		Sigmoid colon	01.09.1997		172	70
3B	0	1		Aterosclerotic heart disease	06.03.1999		170	62
3B	0	X		Caecum	12.11.2001		183	80

Use of international classifications: e.g. ICD-N

Remove columns having date type

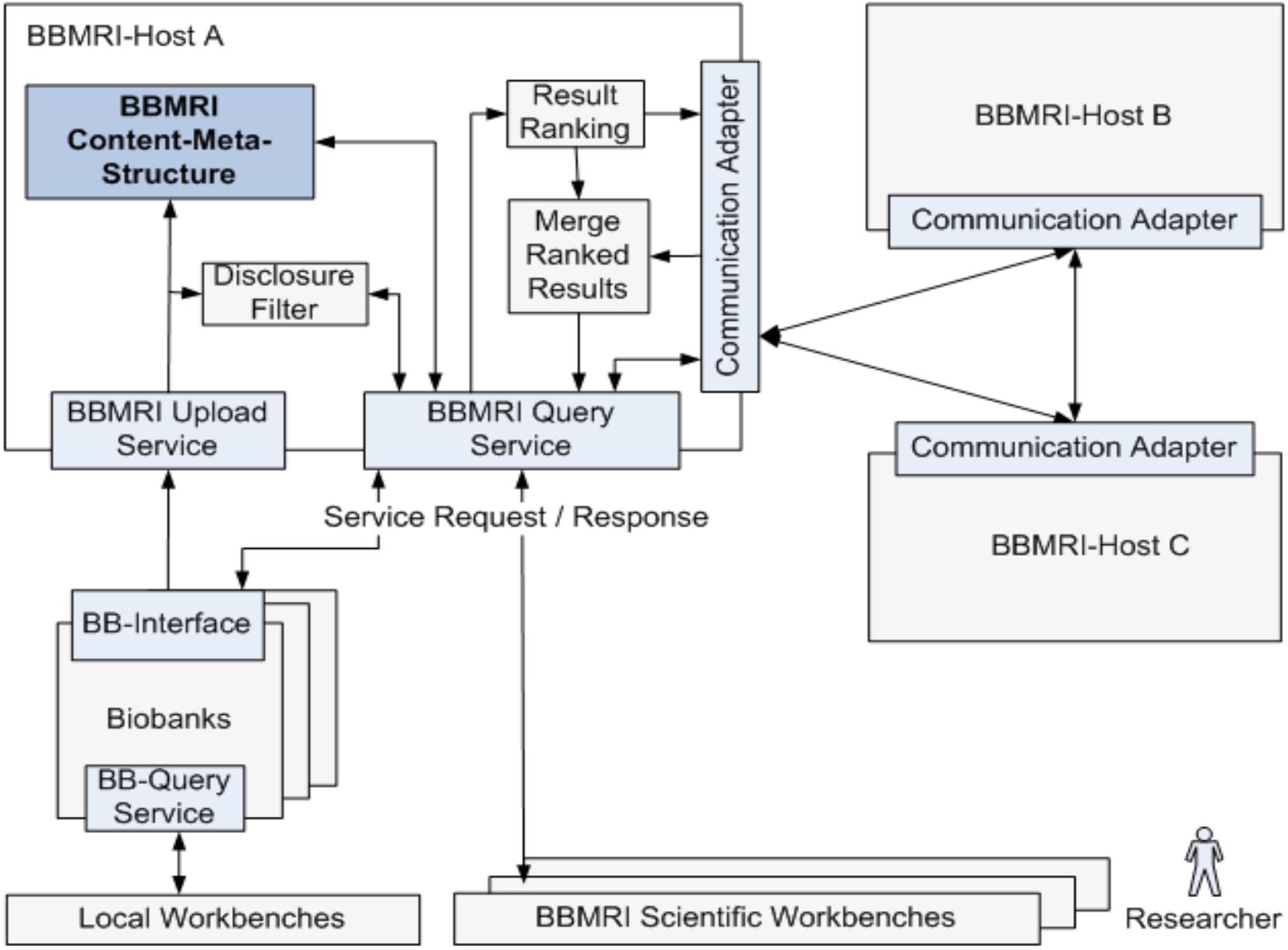
Calculate body mass index

Example: Anonymized data

Staging T	Staging N	Staging M		Cause of Death		BMI
3	0	1		I25		[25.0, .., 29.9] Overweight
3	0	X		C18		[25.0, .., 29.9] Overweight
3	0	X		C18		[18.5, .., 24.9] Normal
3	0	X		C18		[18.5, .., 24.9] Normal
3	0	1		I25		[18.5, .., 24.9] Normal
3	0	X		C18		[18.5, .., 24.9] Normal

Global Integration – BBMRI Project

GEN-AU
GENOMFORSCHUNG IN ÖSTERREICH



Challenges within BBMRI

- Partiality
  - A Biobank as a node in a federation needs descriptive capabilities to be useful for other nodes in the network and it needs the capability to make use of other biobanks.
  - This needs careful design of metadata about the contents of the biobank, the acceptance and interoperability of heterogeneous partner resources.
- Auditability
  - Documenting the origins and the quality of data and specimens, documenting the sources used for studies and the methods and tools and results of studies is essential for the reproducibility of results.
- Longevity
  - Incorporating of changes like: new therapies, new analytical methods, new legal regulations and new IT standards.
- Confidentiality
  - IT-Infrastructure must provide means to protect the confidentiality of protected data and enable the best possible use of data for studies respecting confidentiality constraints.

GATIB II SP4
Research Goals

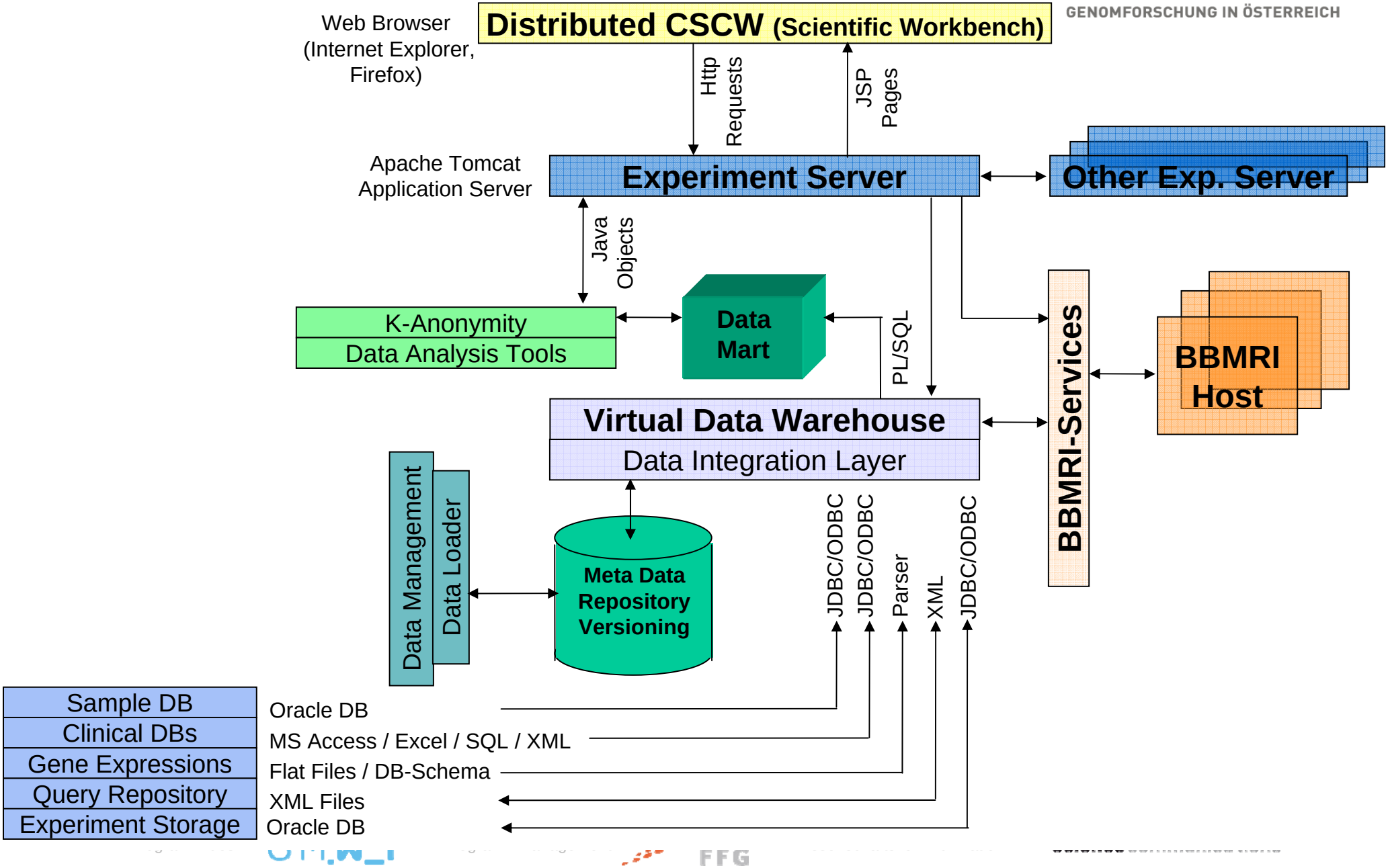
GEN-AU
GENOMFORSCHUNG IN ÖSTERREICH

1. Integration of Biobanks within BBMRI
   - "Federation of European Biobanks" as essential research infrastructure
   - Service based architecture to implement BBMRI interface specifications
   - Integration of clinical databases and data collections (patient data, clinical records, questionnaires, etc.
   - Security and privacy is a key issue

2. Distributed CSCW system for multicenter studies
   - Researchers workbench extended for distributed work
   - Modeling and enforcement of legal/ethical rules for data exchange
   - Coordination for joint projects

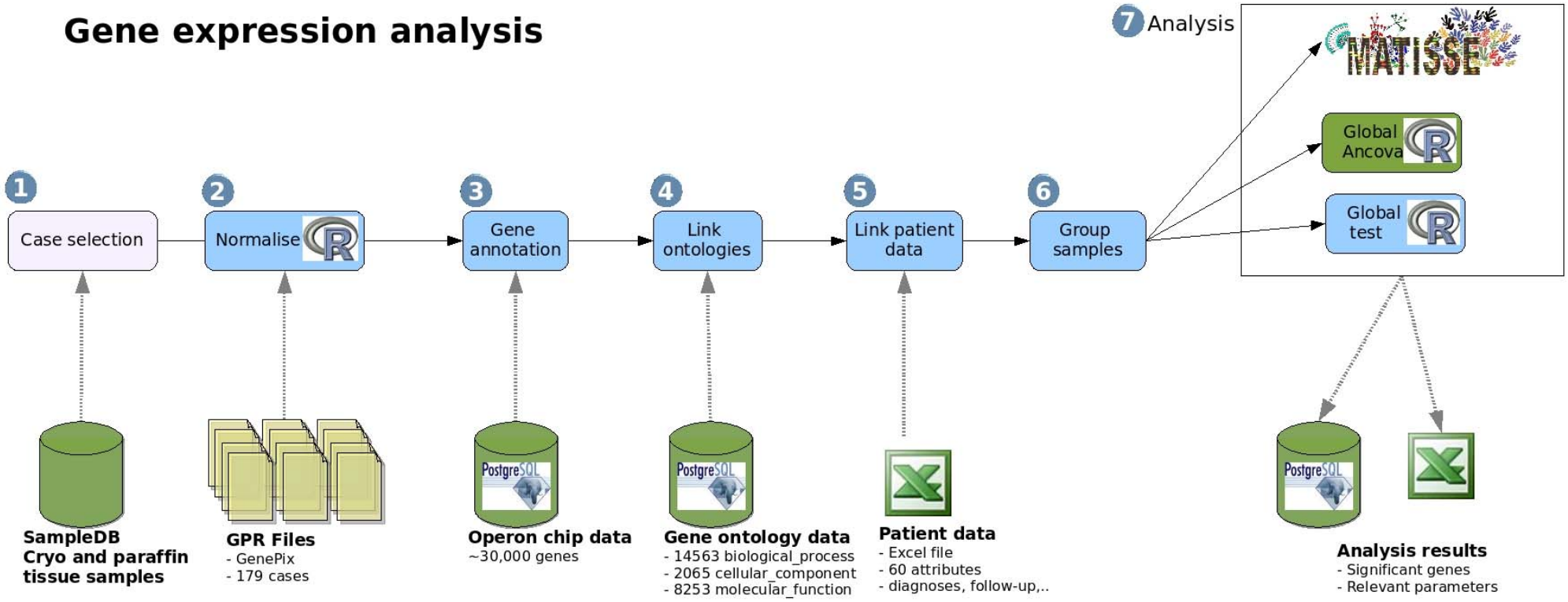
3. Versioning and Evolution of databases
   - "historical databases" to support longevity of data collection
   - Mapping between different versions (e.g. diagnosis codes, etc.)
   - Re-evaluation of results when underlying data (e.g. GeneOntology) changes

IT Architecture



GEN-AU
GENOMFORSCHUNG IN ÖSTERREICH

Example Workflow



Conclusions

- Information system for biobanks
  - Requirements
  - Challenges in the data management
- Integration problems
- Support for medical processes
  - E.g.: Gene Expression Analysis
- Protection of the anonymity of patients
  - Data twins
  - K-anonymity
- Challenges of the global integration
  - BBMRI Project