



Grant Agreement number: 212111

Project acronym: BBMRI

**Project title: Biobanking and Biomolecular Resources Research Infrastructure**

**Project website URL: <http://www.bbmri.eu/>**

**Project Coordinator Name and Organisation:  
KURT ZATLOUKAL, Medical University of Graz (MedUG)**

**E-mail: [kurt.zatloukal@meduni-graz.at](mailto:kurt.zatloukal@meduni-graz.at)**

## PROJECT DELIVERABLE

### D5.6 Final report

**Work Package Leader Name and Organisation:  
JAN-ERIC LITTON, Karolinska Institutet (KI)**

**E-mail: [jan-eric.litton@ki.se](mailto:jan-eric.litton@ki.se)**

---

**Deliverable Due date (and month since project start): 31/01/2010 (m24)**

**Deliverable Version: v1.0**

This page intentionally left blank.

## Document history

Version	Date	Changes	By	Reviewed
0.1	2010-04-14		Martin Fransson	
0.5	2010-05-02		Martin Fransson	Jan-Eric Litton
0.8	2010-05-05		Martin Fransson	Juha Muilu, Paul Fliccek, Jan-Eric Litton, Johann Eder
0.9	2010-07-05	Added Abstract, Updated Section 1.4 Conclusions, Added Section 1.5 with BBMRI-ELIXR statement	Juha Muilu, Martin Fransson	Jan-Eric Litton
0.91	2010-08-16	Linguistic revisions, Minor additions to Abstract, Section 1.1.3 and Section 1.4.2, Updated Section 2 with D5.7, Added Section 4 Acknowledgements	Martin Fransson, Jan-Eric Litton, Juha Muilu	Jan-Eric Litton
1.0	2010-08-25	Minor revisions in Abstract Removed Section for Explanation of the use of the resources, which is to be reported later	Jan-Eric Litton, Martin Fransson	

This document leverages on and integrates the following documents:

Title	Authors	Integrated parts as
D5.1 Inventory of standard related issues.	Martin Fransson ( <i>Karolinska Institutet</i> )	Section 1.1.1
D5.2 Strategy for unique and secure identities for specimens, subjects and biobanks.	Klaus Kuhn S. Wurst D. Schmelcher G. Lamla F. Kohlmayer ( <i>Technical University of Munich</i> )	Section 1.2.1 Section 1.3.4
Unique and secure identities for specimens, subjects and biobanks.	Andy Harris ( <i>UK Biobank</i> )	Figure 2
D5.3 Strategy for communication between biobanks including a common nomenclature, compatible software techniques and appropriate information transmission policies.	Johann Eder Claus Dabringer Michaela Schicho ( <i>University Klagenfurt</i> ) Hákon Gudbjartsson ( <i>deCode Genetics Inc.</i> ) Klaus Kuhn ( <i>Technical University of Munich</i> ) Luciano Milanesi M. Gnocchi D. Ronzoni ( <i>Institute of Biomedical Technologies ITB-CNR</i> ) Jan-Eric Litton Martin Fransson ( <i>Karolinska Institutet</i> )	Section 1.2.2 Section 1.3.1 Section 1.3.2 Section 1.3.3 Section 1.3.6 Section 1.3.7
D5.4 Requirements for a general information management system for European biobanks.	Johann Eder Claus Dabringer Michaela Schicho ( <i>University Klagenfurt</i> )	Section 1.1.2
D5.5 Strategy for a federated hub and spoke structure for European Biobanking.	Juha Muilu ( <i>Biomedicum Helsinki</i> )	Section 1.1.3 Section 1.2.1 Section 1.3.5

This page intentionally left blank.

## Abstract

### **Database Harmonisation and IT-infrastructure,**

Workpackage 5 (WP5) coordinates and supervises all processes of the IT, informatics and infrastructure in BBMRI preparatory phase.

The move towards a universal information infrastructure for biobanking in Europe is directly connected to the issues of semantic interoperability through standardized message formats and controlled terminologies. The BBMRI network is composed of multiple national or local hubs connected together in a federated manner.

The WP5 final report is the joint effort by forty-two persons from eleven countries (See Section 4).

Formally, WP5 has been divided into three tasks:

- Task 1: “Requirements for a general information management system for biobanks in Europe”
- Task 2: “Systems for maintaining unique and secure identities for specimens, subjects and biobanks”
- Task 3: “Strategy for communication between biobanks, including a common nomenclature, compatible software techniques and appropriate information transmission polices”

In relation to these tasks, a federated infrastructure with national or regional hubs and the local biobank databases as main components have been proposed, using three levels of data federation:

1. Meta-data
2. Aggregated data
3. Object data – subject or/and sample data

Architecture should be based on Service Oriented Architecture (SOA) pattern using standard data formats and application programming interfaces (APIs). Implementation should be based on standard web-service and grid technologies. The proposed architecture of database federation has been partially demonstrated by the two prototypes. A proposal for a generalised data model has also been developed as means towards data sharing in BBMRI. The data model is dynamic since each biobank may choose if a particular attribute should be of content- or existence type. The data model is adaptable to different kinds of biobanks by using different kind of schemas. What attributes that should reside in a particular schema (e.g., for cancer biobanks) must be decided by an expert group for the specific domain. The common set of attributes for all study types would define the minimum data set, for which a first version exists.

Data federation for meta-data and aggregated data do not require globally unique identifiers (GUID) issued to be maintained by an external authority. Hence, surrogate identifiers, which should not contain any semantics, should be used. Exclusion of the semantic information from identifiers makes them more stable. It is important that identifiers can be created and managed locally in a coordinated fashion. If need for a system for globally unique identifies should arise, ISO/HL7 OIDs will be a good choice as they are existing in the health care domain

already. Mapping to the surrogates is possible by maintaining 1:1 mapping to surrogate keys, which are managed locally. A final decision on a GUID standard for biological information should be made jointly with other affected ESFRI (The European Strategy Forum on Research Infrastructures) projects.

Work on IT and a Data Protection deliverable is ongoing jointly with WP6, using WP5 derived user scenarios.

This page intentionally left blank.

## Table of Contents

1	Work progress and achievements during the period .....	9
1.1	Task 1: Requirements for a general information management system for biobanks in Europe .....	9
1.1.1	Considerations derived from in-depth interviews .....	9
1.1.2	Use cases and workflow .....	11
1.1.3	Formalized requirements .....	16
1.2	Task 2: Systems for maintaining unique and secure identities for specimens, subjects and biobanks .....	19
1.2.1	Inventory of relevant GUID systems and recommendation .....	19
1.2.2	Architectural considerations on security and privacy .....	20
1.3	Task 3: Strategy for communication between biobanks, including a common nomenclature, compatible software techniques and appropriate information transmission polices .....	22
1.3.1	A generalized metadata model for regional BBMRI hubs .....	22
1.3.2	Prototype A – An early version of the generalized metadata model .....	33
1.3.3	Prototype B – Based on the Set Definition Language (SDL) .....	38
1.3.4	Scenarios for service architecture .....	43
1.3.5	Network and implementation model .....	47
1.3.6	The minimum data set .....	50
1.3.7	The Biobank Lexicon .....	52
1.4	Conclusions .....	53
1.4.1	Task 1: Requirements for a general information management system for biobanks in Europe .....	53
1.4.2	Task 2: Systems for maintaining unique and secure identities for specimens, subjects and biobanks .....	53
1.4.3	Task 3: Strategy for communication between biobanks, including a common nomenclature, compatible software techniques and appropriate information transmission polices .....	54
1.5	External collaboration .....	55
1.5.1	BBMRI/ELIXIR Working Group Statement: 16-17 Nov 2009 .....	55
2	Deliverables and milestones tables .....	57
2.1	Deliverables (excluding the periodic and final reports) .....	57
2.2	Milestones .....	58
3	Acknowledgements .....	59
4	References .....	61



# 1 Work progress and achievements during the period

For each task specified for Work package 5 “Database harmonisation and IT-infrastructure” in the Grant Agreement (GA BBMRI, no.: 212111) the following sections present the outcomes of task-specific results in deliverables, and also discusses any deviations for these from the original work plan.

## 1.1 Task 1: Requirements for a general information management system for biobanks in Europe

*Task 1 in this work package will be to arrive at a consensus on the requirements for a general information management system for biobanks in Europe. (GA BBMRI, no.: 212111)*

Primary related deliverables: D5.1, D5.4, D5.5

At least three of the WP5 deliverables relates to requirements for a system for federation of biobank data. Deliverable D5.1 constitutes an inventory of experiences in IT-systems development and information management in large-scale biobank organizations. D5.4 contains user scenarios and more general use cases from which requirements can be elicited. A workflow is also presented. In D5.5 requirements, including the ones from D5.4, have been further generalized and grouped into different categories. The main results from each deliverable are outlined below.

### 1.1.1 Considerations derived from in-depth interviews

In D5.1 “Inventory of standard related issues” experience collection has been undertaken through a number of in-depth interviews at major biobank initiatives. The outcome of the interviews provides a qualitative measure, complementary to the quantitative measures covered by the IT-supplement of the BBMRI-questionnaire. The experiences can be reformulated as soft requirements or *considerations*, which if adhered to, are likely to facilitate the overall quality of the system and the information content.

#### 1. Considerations on data collection

In addition to SOPs (Standard Operating Procedures) for sample management, testing specific considerations for data collection, e.g., sample information extraction or database input, could be used to give an overall indication of the quality of a particular data set. This measure would precede the inclusion of a new data set into the hub-and-spokes network for data federation. Data sets with several considerations tested as positive could be ranked as being of particularly good quality. The considerations are not strictly defined and the list could be extended.

a. *Have data modelling been considered prior to data collection?*

A predefined data model implies a better structure of the data, and could ensure better consistency in data definitions over time.

b. *Are data defined within a context?*

A context, which may be part of the data model, for the data item may improve quality since the context will describe some of the conditions for which the data value was obtained. This measure will for instance facilitate comparability of numerical values for data items that have the same name across different studies.

- c. *Have automatic data transmission from the measuring device to the computer been used?*

Automatic data transmission may prevent errors from human factors.

- d. *Are automated checks used to validate manually typed data values?*

Automated checks may prevent typing errors, such as misplaced decimals that would result in impossible values.

## 2. **Considerations for a shared data model**

The network of biobanks will constitute a federated system for which a data model should be designed. The data model could be derived from an existing standard or completely developed in-house. The following considerations should be taken into account for the shared data model.

- a. *Capability to deal with changes in data definitions*

- i. *in general*

Biobank data is supposed to persist for a very long time. Hence definitions for data items are likely to change, which means the model structure need to flexible to enough to deal with potential changes. Flexibility in this context could mean keeping the model less complex and trying to design the model so that data entities are relatively independent.

- ii. *in relation to medical ontologies*

A shared data model is likely to use definitions from standard medical ontologies such as ICD-9, ICD-10 or SNOMED CT. For this reason the shared data model will also be dependent on the versions of these different standards. Biobanks participating in the network that are using the shared data model may have different preferences for different standards. This issue should be taken into account when designing the model. Preferably, an umbrella ontology, like the UMLS (Unified Medical Language System), should be used for mappings between different standards.

- b. *Separation of phenotype and genotype data*

Existing clinical data models and in-house developed models tend to focus on diagnosis and phenotype information. However, there is also an increasing amount of genotype data in research and medical care. It is likely that this type of data will be requested as part of the shared data model, at least in a later stage. To plan for how such data can be incorporated, already at the first design of the shared data model, could make it easier to include the genotype data in a post-hoc manner. It is plausible that a distinct separation of phenotype and genotype data entities will facilitate future extensions for the shared data model.

- c. *Natural language perspective*

BBMRI is a pan-European network, which means many countries and languages are involved. Participating biobanks may prefer to have data definitions in their own national language (like medical records). However, a shared data model is likely to be developed using a reference language, and most likely English. To facilitate use of the data model in non-English speaking countries BBMRI should make sure that the data definitions are also available in the languages of the participating nations.

### 1.1.2 Use cases and workflow

D5.4 “Requirements for a general information management system for European biobanks” aims at describing requirements for BBMRI informatics from an end-user perspective. Requirements are derived from a detailed user-scenario and more generalized use cases. Based on the use cases a workflow is derived.

The use cases are needed to determine the requirements for a BBMRI information management system. Further, they can be used for identifying the capabilities of BBMRI now and in the future. WP5 meeting discussions and reports have resulted in a list of use cases. The list starts with the most basic use case and evolves into more complex ones. Since a particular use-case class is dependent on the implementation of the less complex preceding class the list is also automatically in priority, with highest priority for the first use case. The higher the complexity of a use case is the lower its priority.

1. *Search for biobanks*. Retrieves a list with contact data from participating biobanks that have desired material for a certain study. This first class is sub divided into three more classes:
  - 1a. *Distributed metadata queries*. Search for availability of attributes.
  - 1b. *Distributed sample counts*. Retrieves approximately amount of available samples.
  - 1c. *Search for detailed data for samples and subjects*. Operates on local databases of the participating biobanks and retrieves a list with contact data from the biobanks.
2. *Search for cases*. Retrieves the pseudonym identifiers of cases stored in biobanks that correspond to a given set of parameters. In our context, a case is set of jointly harvested samples. This use case is sub divided into two classes
  - 2a. *Distributed metadata queries*. Operates only on the metadatabase and returns pseudonym case identifiers.
  - 2b. *Search for detailed data for cases*. Operates additional on the local databases of the participating biobanks and retrieves only a pseudonym unique identifier of the appropriate cases.
3. *Statistical queries*. Performs analytical queries on a  $k$ -anonym dataset [1] of biobanks.
4. *Retrieval of detailed data*. Obtains available information (material, data, etc.) for a given set of parameters directly from a biobank. The obtained information can then be used for further downstream analysis or in a federated processing.
5. *Upload or linking of data*. Connecting samples with data generated from this sample internally and externally.

Figure 1 shows a possible workflow for the search for biobanks and cases, separated into different responsibility parts. The most important participants within this workflow are the requestor (a registered and authenticated researcher), the requestor's BBMRI host, other BBMRI hosts and biobanks. Hosts act as global coordinators within the federation. The registration of biobanks on BBMRI hosts takes place via a hub and spoke structure. The clear separation of the workflow into different responsibility parts helps when identifying the needed interfaces of each subsystem.

In the following each step of the workflow is summarized and a description of the produced output is given. The output of each step is passed as input to the succeeding step.

### S1 - Select Service Request:

In the first step of the workflow an authenticated researcher can choose a service request from a list of available services. The list of available service requests is meant to support the researchers by providing predefined requests for common queries. The predefined service requests do not limit the system. It is possible to post arbitrary queries on the metadata. Since a request on material or medical data can have different conditions, a suggestion is to provide a list of possible query templates like:

- Biobanks with diseased samples (cancer)
- Biobanks with diseased samples (metabolic)
- Cases with behavioural progression of a specific kind of tumour
- Cases with commonalities of two or more tumours
- ...

**Output description:** Selected service identified by unique ID and name.

**Output type:** XML file

S2 – Define Filter Criteria: After the selection of an appropriate service request the researcher can declare service-specific filter criteria to constrain the result according to the needs. Since BBMRI has to deal with a federated heterogeneous set of different databases we have to use approximate query answering techniques. The researcher's possibility to specify a level of importance for each filter criteria helps in dealing with the approximate query answering. This level of importance is an interval between 1 and 5 with 1-lowest relevance and 5-highest relevance. Without any specification, the importance of the filter criteria is treated as default-value 3-relevant. The level of importance has direct effects on the output of the query. It is used for three major purposes:

- *Specifying must-have values for the result.* If the requestor defines the highest level of importance for an attribute, the query only returns databases that match exactly.  
*E.g.:* Searching for biobanks that must store *Diagnose, PatientSex and Gender* will not return a database that only stores a subset of those three attributes.
- *Specifying nice-to-have values for the result.* This feature relaxes query formulations in order to incorporate the aspect of semi-structured data.  
*E.g.:* Searching for biobanks that eventually store *Diagnose, PatientSex and Gender* will also return a database that only stores a subset of those three attributes.
- *Ranking the result to show the best matches at the topmost position.* The ranking algorithm takes the resulting data of the query invocation process and sorts the output according to the predefined levels of importance.

All these filter criteria together with the level of importance form the so-called importance template.

**Output description:** Individual specified filter criteria based on selected service request inclusive level of importance for each criteria.

**Output type:** XML file as importance template

Researchers formulate their requests with the use of query by example. From the researcher's perspective, BBMRI acts as one single system, performing query processing and disclosure of information from the participating biobanks transparent. According to this, the formulated query of the researcher is sent to the requestor's national host as XML document.

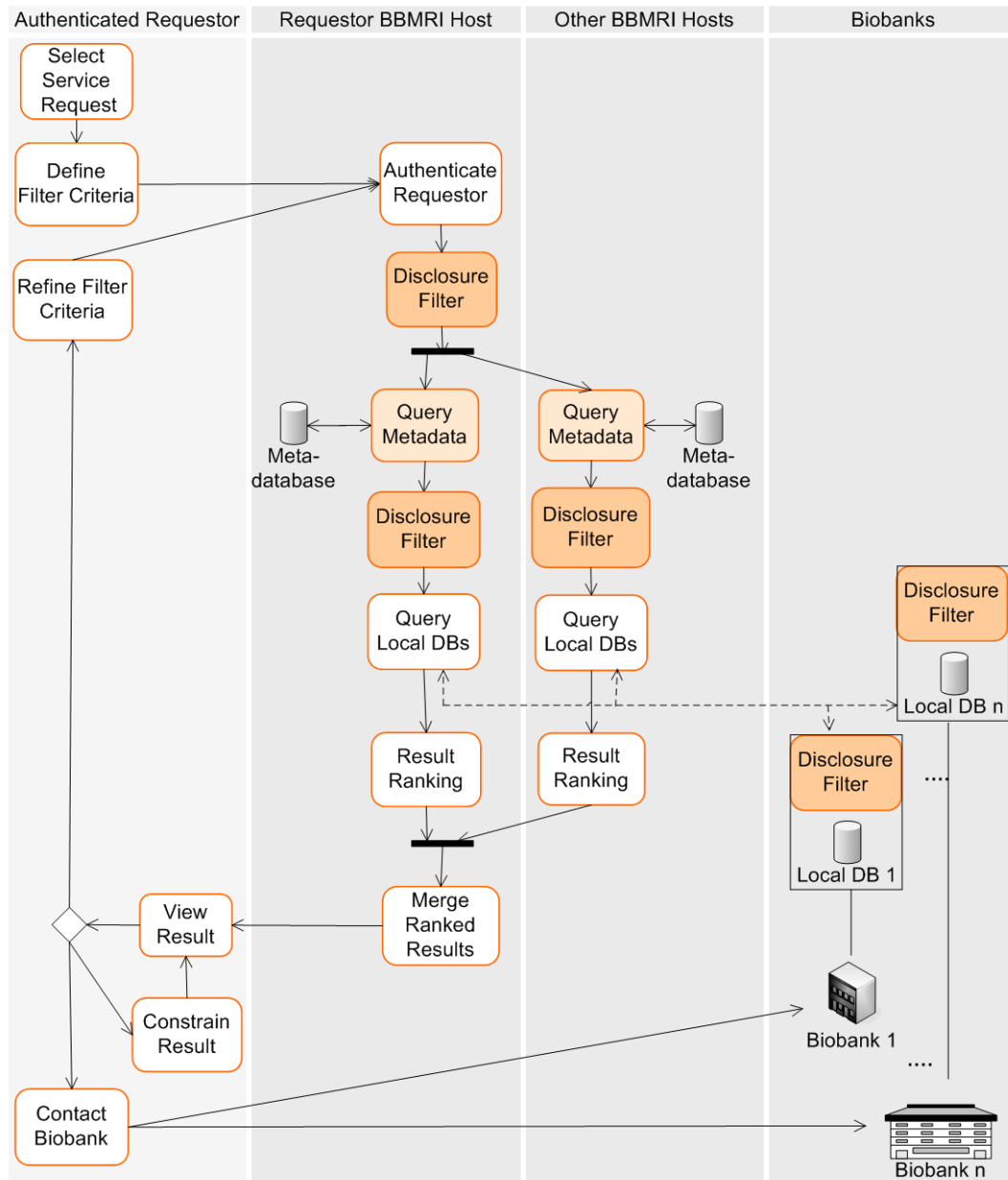


Figure 1: Workflow for a use case separated into different responsibility parts

**S3 – Authenticate Requestor:** User, role and rights management plays an important role within BBMRI. At the beginning of using BBMRI, each user must prove one's identity. After the registration step, the new user receives a username and an initial password that must be changed immediately after the first login. The username/password tuple must be used each time accessing BBMRI. Within the workflow, there is one-step that authenticates the requestor, i.e., checks username and password of the requestor.

S4 – Disclosure Filter: The disclosure filter is a software component that helps the BBMRI-Hosts to answer the following question: *Who is allowed to receive what from whom under which circumstances?* E.g., since it is planned to provide information exchange across national borders the disclosure filter has to ensure that no data (physical or electronic) leaves the country illegally. The disclosure filter should take into account laws, contracts between participants, policies of participants and even rulings (e.g., by courts and ethics boards).

The disclosure filter on a BBMRI-Host plays three different roles:

1. Provider-host and local biobank remove items from query answers that are not supposed to be seen by requestors.
2. Technical Optimization: Query system optimizes query processing using disclosure information.
3. Requestor-host removes providers that do not provide sufficient information to the requestor. This role can be switched on/off.

After applying the disclosure filter, the national host distributes the query to the other participating hosts in the federation using disclosure information. Each invoked BBMRI-Host in the federation performs the same procedure as the national host, but without distributing the incoming query.

S5 – Query Metadata: The national host and each BBMRI-Host, which received the distributed service request, queries its own metadatabase. Depending on the service request and disclosure information additionally also the local databases of the actual registered biobanks at that hosts are accessed. This only happens in use case 1c and 2 and is therefore illustrated with a dashed line in Figure 1.

**Output description:** List of biobanks (name, contact information, requested items) or additional a list of cases, identified by unique ID, which hold the requested data or material based on the specified attributes.

**Output type:** XML file

S6 – Result Ranking: Since it is very useful to the requestor to find the most important result entries at the first output lines an intelligent ranking module is provided. The ranking module takes the resulting data of the query invocation process and sorts the output according to a predefined importance template, which has been generated from the data given in *S2 – Define Filter Criteria*. The importance template is written in XML and contains definitions of the importance of available attributes and values. The more likely the result and the importance template are, the more important the result is treated – thus leading to a higher ranking. Additionally to the importance of attributes, also the density of available values in the local databases is an operative aspect of the result ranking. The more likely the result and the requested order of magnitude are, the more important the result is treated.

After the parallel execution of the distributed ranking modules, the ranked lists are passed to the module that merges the several ranking results.

**Output description:** List of biobanks (name, contact information, requested items, matching index) resp. identified cases, which hold the requested data or material based on the specified attributes. This list is ranked by a matching index that represents the likelihood of importance

template and query result. The matching index is a percentage factor and can therefore take any number between 0 and 100.

**Output type:** XML file

All distributed ranked query results are sent back to the requestor's host and are merged on it.

*S7 – Merge Ranked Results:* This step takes several XML files, containing the ranked list of biobank information, and merges them to one single XML file. This XML file finally contains all ethical authorized biobanks of the BBMRI host and all other participating country hosts that satisfy the filter criteria in a ranked order (best match first). Because the ranking algorithm is executed in a distributed manner, it must be ensured that all BBMRI hosts use the same procedure to rank their results. Otherwise, the merging process would lead to inaccurate or even wrong results.

*S8 - View Result:* A list of biobanks or a list of biobanks including identified cases, which have the requested data, is displayed. To each biobank in the result all available attributes resp. items are shown. At this step the requestor can choose between three different ways to proceed:

- *Refine Filter criteria.* In case of a too large or too small result the requestor may want to refine the given filter criteria or importance of step S2. As already described in S2 the requestor has the opportunity to set filter criteria and importance weights for each of the attributes. After the refinement process is finished, the ethical filter has to ensure the authorization of the request within the participating federated biobanks. This step leads to a re-execution of the query in the distributed system of BBMRI hosts and biobanks.
- *Constrain current result.* Choosing this step gives the requestor the possibility to constrain the result set in terms of attributes that are part of the result. Since the attributes are already part of the result, this step does not lead to a re-execution of the query. The following attributes can be constrained by the requestor:
  - *Name of biobanks:* The requestor can constrain the result set for biobanks with a given name – wildcards are supported.
  - *Contact information:* The requestor can constrain the result set for biobanks in specific countries, territories, regions and so on...
  - *Matching index:* The requestor can specify a range for the ranking index to occur in the result. For instance, the requestor only wants to view results matching between 80 and 100 percent.
- *Contact one or more desired biobanks.* The requestor can choose one or more desired biobanks and get in contact with them. The contact information can be the complete postal address, email or even a phone contact. For computer automated contacting of the biobank an URL can also be given.

### 1.1.3 Formalized requirements

D5.5 “Strategy for a federated hub and spoke structure for European Biobanking” aims at outline an implementation strategy for a hub and spokes network. To facilitate a set of implementation principles the report first describes formalized and categorized requirements from WP5 discussions and previous deliverables.

#### 1. Technical requirements on BBMRI data integration system

The following general requirements have been identified:

- R1. Given the complexity, usage of the information system should be divided at least into two complementing parts addressing
  - 1.1. Needs of resource (here mainly data and samples) discovery and
  - 1.2. Sharing of original data for research purposes.
- R2. Local database polices, national ELSI regulations and EU data protection act must be followed.
- R3. The identification scheme for samples and subjects can be based on surrogate identifiers maintained by co-operating systems providing context for the identifiers. Global identification scheme is necessary if identifiers are taken outside the context.
- R4. Sample and subject identifiers must be randomized and identifiers should not contain any meaning.
- R5. Each biobank must identify their specimen and related information adequately and persistently.
- R6. The data integration framework must have possibility for data federation without sacrificing benefits of centralized approaches where data is collected into one single database.
- R7. Standard security protocols and measures must be used. There are at least two different security domains related to data discovery (R1-1.1) and data analysis (R1-1.2) having different security requirements due to nature of data.
- R8. All queries and/or access to data services and analysis tools should be logged and data provenance issues taken into account. The auditing information should be stored for determined time.
- R9. Users must be authenticated and authorized, e.g., via federated architecture such as OpenID [2]. Each country must be able to register and manage the credentials of local users.
- R10. Authentication and authorization should be done on a level (like in a local hub) where identification of users is most reliable. A central repository can be used to support access control in cases of possible policy violations.



- R11. Database systems must be kept up-to-date. Access to monthly archives should be accessible at least 10-15 year back in time.
- R12. Core informatics needs related to the hub-and-spokes network are same in all participating network nodes (hubs and biobanks). Developed applications and software can benefit all nodes. Common core needs must be defined in design phase.
- R13. Application programming interfaces (APIs), data formats and vocabularies must be standardized. Existing standards must be used wherever possible.
- R14. Local biobanks should have control on the data they expose (“the mine problem”).

## 2. Special data federation requirements

One of the key factors in data integration of distributed systems is the extent of data localization, i.e., how much data is cached or stored outside source databases. Different integration scenarios are presented in Section 1.3.4. The following influencing requirements have been identified:

- R15. Only  $k$ -anonymized [1] data and metadata is allowed to leave each biobank node without explicit permission for down-loading detailed data (R2).
- R16. Data processing should be distributed to the source biobank nodes and no identifiable information should leave each node.
- R17. System should have sufficient level of redundancy for minimizing system downtime and increasing data transfer bandwidth.
- R18. Level of independency: National or local networks must be functionally independent from the parent network, i.e., local services should not be hampered by external factors meaning that at least national (or local) metadata must be stored into national (local) hub.
- R19. Data access use cases must be implementable. For analysis purposes it can be essential to collect relevant data into one place. Data sets can be processed faster and kept stable.
- R20. Distribution of data management and curation work. Curation of primary or derived data should be done on sites having knowledge and expertise on the data.
- R21. Metadata should be defined in a way that it can be collected into centralized data marts.
- R22. Metadata for which  $k$ -anonymity cannot be guaranteed must not be collected outside biobanks (or possibly outside local hubs).

## 3. Networking requirements

The hub-and-spokes model has been proposed as a basic unit for the data integration architecture because of its simplicity and scalability. In the model, network connections (spokes) are arranged so that all traffic from connected nodes goes through a central hub working as a message broker. A drawback of the approach is

that a hub presents a single point of failure. Also, communication can be slow because of the extra step posed by the hub. These problems can be addressed:

- R23. Increasing redundancy of the system so that hubs can take responsibilities from others. Also services and network connections must be monitored constantly.
- R24. Alternative network strategies, like peer-to-peer connections, should be allowed. This is especially important when transferring actual data since volumes can be huge compared to the metadata.

#### 4. Data schema and access requirements

- R25. The BBMRI network should be shared nothing, meaning that all data that is used to search for any given subject must reside in a single node (biobank and hub). This means that all data derived from samples must "come back home".
- R26. Data access (case R1-1.2) can include manual or semi automated steps where human invention is needed to judge data access and usage rights (called as *disclosure filters* in the architecture model).
- R27. Query and analysis tools should be metadata driven and not custom-written against a fixed data model.
- R28. Query tools and database schemas should support hierarchical and DAG structured vocabularies and ontologies.
- R29. Users must be able to specify dynamically the attributes and the scope of the database that is used in analysis, e.g., aggregation analysis.
- R30. Data schema must support event-based data and query tools should support time-based longitudinal analysis.
- R31. Metadata can be separated based on content, number of cases and existence attributes. Existence attributes can be divided further into *or*-connected quantities and *and*-connected availabilities [3].
- R32. A domain lexicon for biobank informatics must be defined.

## 1.2 Task 2: Systems for maintaining unique and secure identities for specimens, subjects and biobanks

Task 2 will be to explore systems for maintaining unique and secure identities (object models) for specimens, subjects and biobanks, as well as for keeping track of the handling of permissions for use, analytical results and statistical output. Meta-information on quality of specimens and phenotypes will be integrated. (GA BBMRI, no.: 212111)

Primary related deliverables: D5.2, D5.5, Data Protection Deliverable

Task 2 can be split into two parts; the part for which it was originally intended, primarily exploration of systems for Globally Unique Identifiers (GUIDs), and the part for which an additional deliverable was created, the issue of data protection and privacy. For the latter part work is still ongoing in collaboration with WP6 and the final conclusions will be accounted for elsewhere.

Deliverable 5.2, dedicated to the first part of Task 2, does not contain a final decision for a particular existing system for Globally Unique Identifiers (GUIDs). Instead, it outlines the scenarios of what should be made a preceding decision, the one of suitable service architectures for BBMRI in the short and long-term perspective. The service architecture scenarios are connected to the use cases and system design and therefore presented under Task 3 in Section 1.3.4. Additionally, D5.2 contains an inventory of the most important relevant existing GUIDs systems, presented below.

### 1.2.1 Inventory of relevant GUID systems and recommendation

The **ISO Object Identifier** (OID) can be characterized as follows: “An OID is a globally unique string representing an ISO (International Organization for Standardization) [4] identifier in a form that consists only of numbers and dots (e.g., "2.16.840.1.113883.3.1"). According to ISO, OIDs are paths in a tree structure, with the left-most number representing the root and the right-most number representing a leaf. Each branch under the root corresponds to an assigning authority. Each of these assigning authorities may, in turn, designate its own set of assigning authorities that work under its auspices, and so on down the line. Eventually, one of these authorities assigns a unique (to it as an assigning authority) number that corresponds to a leaf node on the tree. The leaf may represent an assigning authority (in which case the root OID identifies the authority), or an instance of an object. An assigning authority owns a namespace, consisting of its sub-tree” [5].

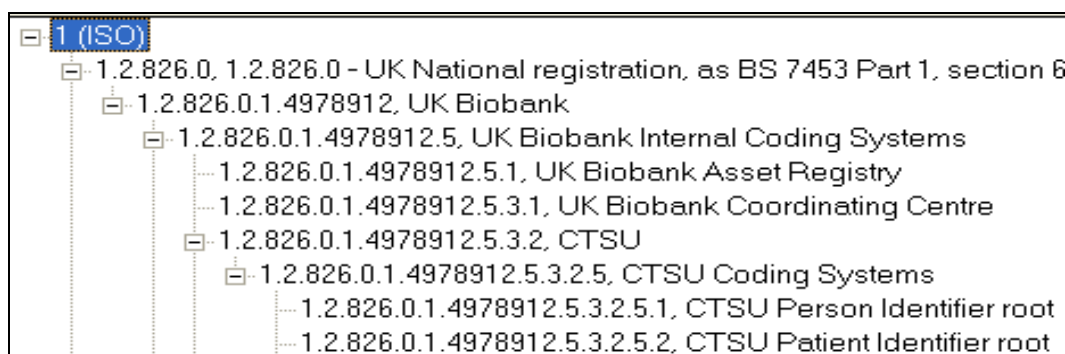


Figure 2: An example of the hierarchical structure of OIDs.

The **Digital Object Identifier** (DOI®) System [6] is a managed system for persistent identification of content on digital networks. It can be used to identify physical, digital, or abstract entities; these names resolve to data specified by the registrant, and use an extensible metadata model to associate descriptive and other elements of data with the DOI Name. The DOI System is implemented through a federation of Registration Agencies, under policies and common infrastructure provided by the International DOI Foundation that developed and controls the system.

**Life Science Identifiers** (LSIDs) [7] are an option to build persistent, location-independent, resource identifiers for uniquely naming biologically significant resources. Different from OIDs, LSIDs are expressed as a URN namespace.

### **Recommendation**

Use cases 1 and 2 in Section 1.1.2 do not require globally unique identifiers issued and to be maintained by an external authority. If a need for such an identifier system should arise, ISO/HL7 OIDs will be a good choice, as they exist in the health care domain already and mapping to the surrogates is possible (D5.2).

Along with the formalized requirements in Section 1.1.3, this recommendation has been formulated as design principal P5 in D5.5:

- P5. Surrogate identifiers, which should not contain any semantics, should be used (R3, R4). Exclusion of the semantic information from identifiers makes them more stable. It is important that identifiers can be created and managed locally in a coordinated fashion. Globally unique identifiers can be made if needed using OIDs maintaining 1:1 mapping to surrogate keys, which are managed locally. Researcher identification and user identification in general is another important issue and it is pioneered in the GEN2PHEN program (EU#200754), where mechanisms for identification, authorization and micro attribution are being developed [8].

In addition to the recommendation above, it has also been in common agreement in WP5 that a final decision on a GUID standard for biological information should be made jointly with other affected ESFRI.

### *1.2.2 Architectural considerations on security and privacy*

Relating to the second part of Task 2 is the design of the federated architecture, primarily discussed in Section 1.3.1, for which the following considerations has been made.

One of the most important issues within the design for the federated architecture of biobanks within BBMRI was to provide security and privacy at the highest possible level. The system was designed in a way that it is impossible for intruders to figure out the identity of a certain donor. It has been shown [9, 10] that the greatest concern of donors is their anonymity and privacy. This fact led us through the whole design process of the federated architecture. The most important design decisions are listed below:

- We do not store any patient related or patient identifying data outside the local biobanks. In the case that the optional extension of storing subject identifiers in the content meta

structure is used we propose using pseudonym identifiers that cannot be tracked back by anyone else except the local biobank.

- The information stored on each hub is  $k$ -anonym and  $l$ -diverse as a whole to prevent trackers from identifying certain donors. It is very important to mention here that also *data increments* must meet the requirements of  $k$ -anonymity and  $l$ -diversity. In the case that newly added items are not  $k$ -anonymous and  $l$ -diverse the data stored within the meta structure can be used to identify certain donors. Further information can be found in [11, 12].
- We assume that *all* involved biobanks implement mechanisms that ensure privacy for all query requests from BBMRI. One possibility to ensure privacy is that queries work on  $k$ -anonymous views. Further, tracker control mechanisms must be used to guarantee security and privacy. Additionally the disclosure filter can be a very helpful tool for supporting these mechanisms.
- In the current scenarios and use cases, no privacy comprising information is exchanged within the BBMRI network. The idea is to find and locate biobanks with promising data or material and delegate the actual exchange of data and material to bilateral procedures, enforcing the local legal, ethical and organizational regulations of the participants.

### **1.3 Task 3: Strategy for communication between biobanks, including a common nomenclature, compatible software techniques and appropriate information transmission policies**

*Task 3 will be to explore a complete strategy for communication between biobank including a common nomenclature, compatible software techniques and appropriate information transmission policies. This all relates to information on specimens, laboratory results, phenotypes, exposures and genealogical data.*

Primary related deliverables: D5.2, D5.3, D5.5

Most work within WP5 has been concerned with Task 3 since this is the most extensive task. It is also dependent on the work performed in Tasks 1 and 2; primarily use cases 1 and 2 in Section 1.1.2. Major activities for Task 3 have been the creation of a shared data model for European biobanks. Two proposals for the design and architecture of an information management system for European biobanks have been developed. However, since the proposals focus on different layers of software technology – web services vs. the Set Definition Language (SDL) [13], they can also be considered to be complementary. In fact, it was suggested that the data schema (presented in D5.3) used for Prototype B in Section 1.3.3 could be viewed as an instance of the generalized metadata model discussed in Section 1.3.1.

Part of Task 3 is also the work related to different service scenarios in Section 1.3.4 from D5.2, and the network model and implementation proposal from D5.5 presented in Section 1.3.5. The minimum data set presented in Section 1.3.6 should only be considered as a draft and an intermediate step for metadata collection before employing a relational model. The first implementation of an online version of a Biobank Lexicon, presented in Section 1.3.7 is a first step towards requirement R32. The master language English has so far been translated to six other languages; Estonian, Finnish, German, Italian, Spanish and Swedish. The Biobank Lexicon is published on <http://www.biobank-lexicon.org>.

#### **1.3.1 A generalized metadata model for regional BBMRI hubs**

A metadatabase is located on each hub in the federated system. This metadatabase stores information of registered biobanks and allows certain queries for researchers. This section deals with the metadata model used for maintaining information on each regional hub. This data model is able to hold information about:

- Relation between Participant databases, Hosts and Participants
- Content stored by certain biobanks – Content meta structure
- Users that may access biobanks with their roles and attached operations.

The stored information helps answering queries as described in use case 1 and use case 2. Further on this information allows distinguishing between different users with different access rights. This feature is needed to ensure that administrators or local biobank users only update data that belongs to their local biobank. In the following we are going to discuss the different areas of the data model shown in Figure 3.

The design rationale for the schema of the BBMRI hubs was to accept that the schemas of the participating biobanks are very heterogeneous. The hubs should support any kind of biobank schema and allow queries against it. On the other hand, we envision a movement towards harmonization and standardization. Our expectation is that such standardization will start in

several research communities (cancer, obesity, etc.) but will be hard to extend to whole large biobanks. Furthermore we have to take into account the dynamics of the field: new types of analysis will appear which will be represented by new attributes in biobank schemas. Such new measures cannot be expected to be established everywhere synchronously, but will first appear in specialized collections and then probably spread out. So even if there could be a homogeneous schema for biobank, the dynamics of the field and the necessity to incorporate new scientific findings and new technological possibilities immediately to support scientific research will make a sustainable standard impossible.

Therefore, we support the definition of standards and the evolution of standards via a concept called *StudyType*, which documents such standards and supports querying. We regard StudyTypes as a loose kind of schema definition, which defines which attributes have to be present in biobank to belong to this StudyType. However, we do not restrict the biobanks in how these attributes are combined in their actual schemas.

The concept *Measurements* supports a very flexible way of representing different schemas.

Querying this structure is not trivial. However, it offers the possibility to support different kinds of query tools and interfaces. For an example starting with study types we can support query tools, which require a rather homogeneous distributed database. On the other hand, starting with arbitrary attributes we can support any kind of semi-structured query interface.

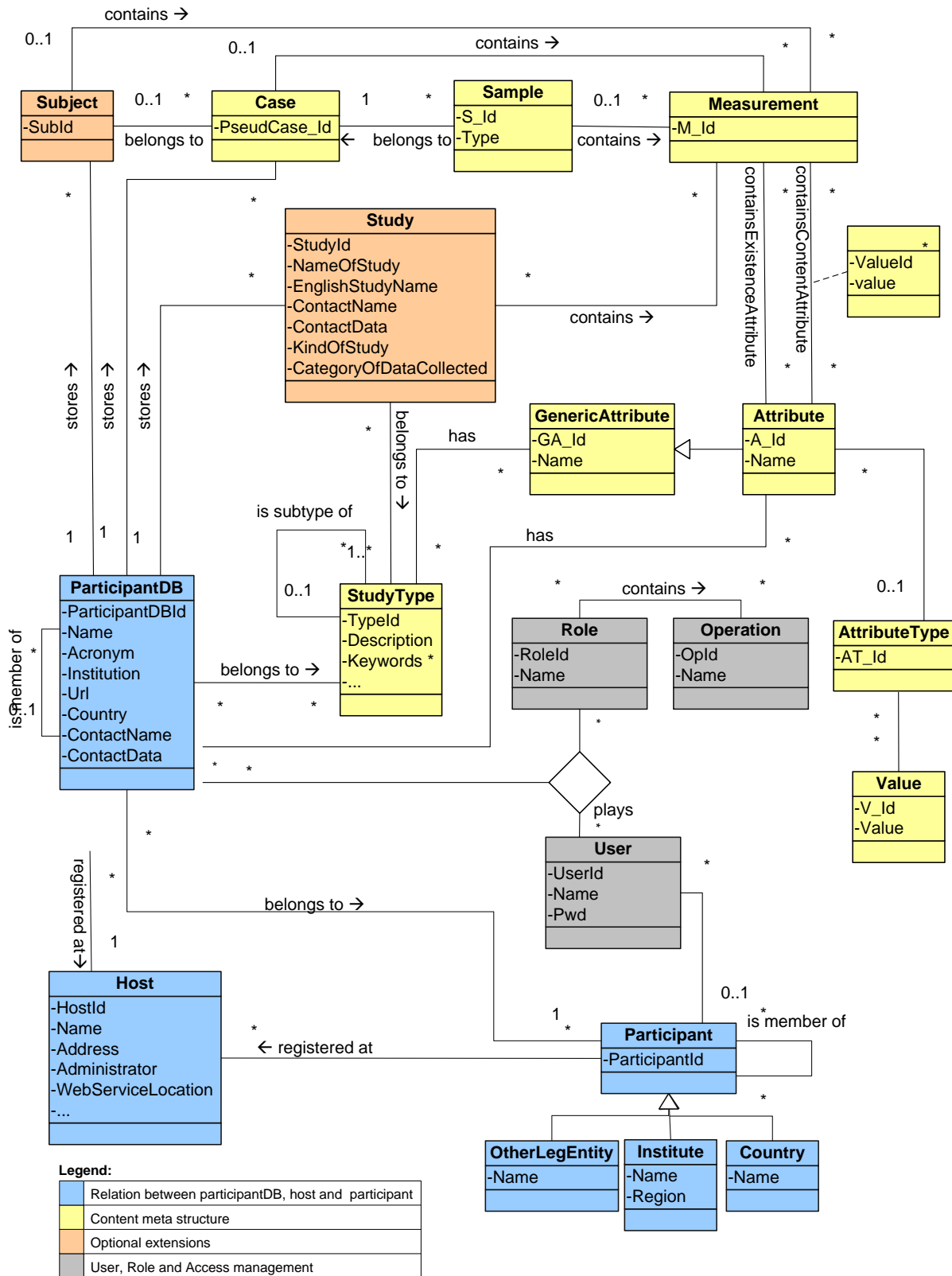


Figure 3: Structure of metadata model for a regional BBMRI hub



### 1.3.1.1 Relation between Participant database, Host, and Participant

In the following section different aspects of the relations between participant database, host, and participant are discussed. It simulates the blue part of Figure 3.

#### Participant database (ParticipantDB)

A *participant database* represents a biobank in the federation. The relation *ParticipantDB* in Figure 3 is used to store the common information, primarily the contact data, about a participating biobank. Participant's databases can be grouped hierarchically and must belong to a participant (e.g.: Institute...).

#### Host

A *host* represents a regional hub in the federation. The relation *Host* holds common information (e.g. location of WebServices, Administrator...) about the particular regional hubs in the federated system. On one hand the class *Host* allows the attaching of an arbitrary number of participant databases to a given host (hub). On the other hand each *Host* is registered at one or more participants (e.g.: Austria – represents the national hub of Austria, EU – represents the regional hub of the European union...) within the federated system.

#### Participant

A *participant* in the federation can either be a country, institute or even another legal entity (like EU, Benelux...). The schema also gives the possibility to group *Participants* hierarchically. This grouping can be done by using the *Participant-isMemberOf-Participant* relation.

### 1.3.1.2 Content stored by certain biobanks – Content meta structure

To boost query answering and reduce the overhead of local querying we suggest the so-called *Content meta structure* (yellow and orange part of Figure 3). The aim of the content meta structure is the ability to hold a sufficient set of needed information structures and schema elements. The idea is that obtainable schema elements/attributes (e.g. export schema) from local databases of biobanks can be mapped with the *BBMRI Content-Meta structure* in order to provide a federated knowledge base for life-science-research.

#### Measurements

The concept *Measurement* supports a very flexible way of representing different schemas. A measurement can be used to *group related* attributes together and connect those grouped attributes to important concepts in the meta-schema. The connection of different attributes can be very versatile. Therefore the concept measurement is connected to the three *main concepts* that should be able to support storing of related attributes. In the proposed schema these three concepts are:

- *Subject*: Attaching certain measurement(s) to a subject allows integrating *population based* biobanks. It is possible to store the *height, weight, Gender, day of birth*, etc together and connect this measurement to a certain subject. Since the mentioned attributes are not part of the schema it is very easy to extend a certain measurement by just adding one more attribute to it. E.g., we could also be interested in storing the *day of birth* of a certain subject. This would not lead to changes in the schema.
- *Case*: Attaching measurements to a case gives the possibility to store related data even if the connected biobank itself is not subject oriented. E.g., we could store the

temperature curve, the pulse rate etc. of a certain case. The meta-schema allows storing multiple measurements for one single case. Adding a timestamp (the granularity is up to the user) to each measurement would enable time-series analysis over different measurements for a case.

- *Sample:* Attaching measurements to a certain sample allows storing information related to that sample. E.g., this supports storing the size, sample type or even multiple diagnoses of one single sample.

In summary: The concept of measurements

- helps us dealing with certain different schemata
- is flexible since introducing new attributes does not lead to changing the schema of a regional hub
- allows grouping an arbitrary number of attributes and connecting them to subjects, cases and samples. E.g., connecting age, Gender and temperature curve to a subject or a case.
- enables attaching multiple occurrences of an attribute to subjects, cases and samples. E.g. it allows storing multiple diagnoses for one sample and also multiple temperature curves for one subject etc.
- is connected via an unbounded relation to subject, case and sample. Therefore it is possible to store multiple measurements for each of those concepts.

## Subject

This optional enhancement of the metadatabase gives the possibility to store a pseudonym identifier for a certain individual and connect either directly to a biobank or to several cases. The identifier must be a pseudonym and only the local biobanks are able to map that pseudonym identifier to a certain person. With storing the pseudonym identifiers of certain individuals we are able to answer questions like:

- Find biobanks with more than 50 subjects that are female and have been measured with very small deviation in their systolic and diastolic blood pressure.
- Find biobanks in Iceland that have data from a study related to obesity and the biobanks must also have more than 100 participants who have BMI attribute (values)
- Find cases of subjects that suffer from
  - breast cancer with available staging TNM and grading
  - lung cancer with available therapy description and a follow up.

As we can see above subjects can be attached to a biobank in three different ways:

- Directly via relation biobank-stores-subject: This relation can be used when a biobank does not store cases nor studies.
- Indirect via Study and Measurement
- Indirect via Cases

## Obtainable schema elements/attributes (Attribute catalogue)

The queries supported by the metadata model are restricted by the available schema elements/attributes (Table 1) provided as so-called *attribute-catalogue* in the content-meta structure. The more attributes are available the more powerful the queries can be expressed. But we must be careful; drawbacks of having too many attributes in the database on each hub

are data overkill and also the problem that not every attribute is available in every connected local database. This fact comes due to the very heterogeneous and autonomous appearance of the local databases.

In order to avoid data overkill the schema contains attributes usually occurring in most or even all of the participating biobanks is recommended. In the following Table 1: a set of schema elements/attributes are shown. This table reflects a rudimentary set of attributes primarily needed for dealing with cancer and population based biobanks and is actually open for extensions to other working areas of biobanks.

Sample specific attributes			Donor / Subject specific attributes		
Type	Quality	Quantity	Common data	Lifestyle data	Measurements
Tissue Blood Urine Cell cultures DNA cDNA/RNA Serum Plasma Fluids	Fresh/Frozen Formalin-fixed Paraffin emb.	Biopsy surgical specimen  few some many	Gender Age at diagnosis AgeGroup Family History Year of Birth Year of Death Phenotype of interest Type of consent <u>Bodysize:</u> BMI Height Weight	Nutrition Physical exercise Alcohol Nicotine Drugs Social status Carcinogen expose	Temperature Bloodpressure <u>Heartfunction</u> Diastolic PulseRate QRS QTC Systolic IsLastMeasure
Anamnesis specific attributes			Cancer specific attributes		Genetic & labor attributes
Therapy description (Chemo, Radiation, Biological) Therapy(Chemo, Radiation) Medicated Follow-up data Clinical data Length of follow-up Disease-free survival Overall survival Cause of death Autopsy OrganCategory Diagnose (ICD-10, ICD O-3, SNOMED CT, Omin) Accessory diagnose Additional diseases SampleDate			Staging TNM (UICC) Grading (Gleeson Grading) Special morphological features Receptor status Receptor type Immunophenotype Mutation status Chromosomal alterations Localisation primarytumor First evidence of metastases Localisation primary metastasis		AssayPlatform Markercount Markername Assayname LOINCCode LOINCComponent LOINCMethod LOINCPROPERTY LOINCScale LOINCSystem LOINCTiming

Table 1: An example outline of obtainable attributes in the metadata model

For the schema of all regional hubs, a universally valid set of relevant schema elements/attributes must be achieved. The so-called *attribute catalogue* acts then as a preconfigured component of the content-meta structure. Ontology for the attribute catalogue provides a description for each of the attributes as well as the management of synonyms. Appropriate transformation algorithms handle different languages and resolve structural conflicts, in order to tackle heterogeneity problems and support semantic mappings within the federation.

In a federation the content (value) of schema elements/attributes is typically not stored in the hub. Due to this fact, for each query that requests for example the specific diagnoses “Liver cancer” the local biobanks of all participating biobanks must be queried. That can possibly result in a high number of local queries. Therefore it is useful that attributes also can have values of content (e.g. Gender: Female, Male) based on an ontology or a thesaurus and have not only an existence status (e.g., is the attribute local available or not). Unfortunately this again leads to rigidity in the federation.

Our approach was to accomplish a hybrid-solution of a federated system and an additional data warehouse as a kind of index to primarily reduce the query overhead, but without rigidity. This decision led to the design of a look-up data mart, which deals with attributes of different meanings, similar to OLAP (online analytical processing), including:

- **Content-attributes.** These attributes allow the storage of information about the content (value) of attributes in the local database to hub. Typically these are attributes with a limitable range of values. These attributes are comparable to dimensions in OLAP.
- **Existence-attributes.** These attributes reflect the occurrences/availability of attributes in a local database for a specific instance tuple of (all) content-attributes. They are comparable to measures in OLAP

The *attribute-catalogue* contains all obtainable attributes, including types and values and is structured in the following way:

- The relation *Generic Attribute* comprises the set of obtainable schema elements/attributes in the metadata model for regional BBMRI hubs.
- The relation *Attribute* is the specialized class of *Generic Attribute* in order to capture different categories of attributes.
- The relation *AttributeType* offers the possibility to configure different types of attributes.
- The relation *Value* holds all available values for a certain attribute type. It is not omitted that for every attribute exist a set of values in the metadata model.

### **The meaning of Study type**

For the specification of different kinds of biobanks the metadata model additionally supports a descriptive *StudyType*-relation. Due to this it is possible to capture different research areas: For instance a metabolic based biobank stores metabolic specific attributes and does not necessarily need to store a TNM-Classification and otherwise a cancer based biobank does not necessarily need to store metabolic specific attributes but cancer specific attributes. That means each biobank can specify its *type(s)* and describe the material and information it harvests.

E.g. *Biobank BBI is a tissue bank, which harvests cryo prepared tissues from the pathology.*

The purpose of the introduced *StudyType* is to get information about:

- Kinds of collected material (Tissue bank, Blood bank...)
- Obtained data (clinical, pathological, ...)
- Kind of disease they work on by the means of the ICD-10 Classification

With such information it is possible to categorise similar types/kinds of biobanks and query over biobanks containing relating attributes is optimized.

The idea is to provide multiple schemas including common sets of specific *generic attributes*, which are significant for particular research areas of biobanks. Each biobank can register their appropriate *StudyType(s)*, by bringing up at least all of the necessary schema elements/attributes. With the relation *ParticipantDB-has-Attributes* the biobanks can then register the actual set of available attributes explicit (in the most cases is it more than in the specified schemas).

We expect that some research communities will develop standards for study types for their particular fields defining all necessary attributes for biobanks to be relevant.

### Example:

In the following tables, we show two exemplary schemas for study types on example of

1. *Cancer based biobank*, see Table 2
2. *Population based biobank*, see Table 3

Sample specific attributes			Donor / Subject specific attributes	
Type	Quality	Quantity	Common data	Lifestyle data
Tissue	Fresh/Frozen Paraffin emb.	Biopsy	Gender Age at diagnosis	BMI, Nutrition Physical exercise Alcohol Nicotine Drugs Social status Carcinogen expose
Anamnesis specific attributes			Cancer specific attributes	
Therapy description Follow-up Disease-free survival Overall survival OrganCategory Diagnose (ICD-10, ICD O-3, SNOMED CT, Omin)			Staging TNM (UICC) Grading (Gleeson Grading) Immunophenotype Localisation primarytumor First evidence of metastases Localisation primary metastasis	

Table 2: Example schema for study type "cancer based biobank"

Sample specific attributes			Donor / Subject specific attributes	
Type	Quality	Quantity	Common data	Measurements
Tissue Blood Urine Cell cultures DNA RNA Serum	Fresh/Frozen Formalin-fixed Paraffin emb.	Biopsy surgical specimen  few some many	Gender Age at diagnosis Year of Birth Year of Death <u>Bodysize:</u> BMI Height Weight	<u>Heartfunction :</u> Diastolic PulseRate QRS QTC Systolic IsLastMeasure
Anamnesis specific attributes			Genetic & labor attributes	
OrganCategory			AssayPlatform Markercount Marker name Assay name LOINCCode LOINCCComponent LOINCMMethod	

	LOINCPROPERTY LOINCScale LOINCSYSTEM LOINCTIMING
--	---

Table 3: Example schema for study type "population based biobank"

For registering to the study-type "cancer" a biobank therefore must provide the entire required schema elements/attributes captured in Table 2. A biobank can register to more than one study type if they provide all the requested schema elements. If a biobank does not fulfil any schema, then this biobank does not belong to any supported study type. Hence, they only declare their available attributes explicit.

### Studies

The schema has the ability to optionally store studies, but it is also able to work without storing any studies. The studies can be managed by the administrator of a certain biobank and are then available for search. They have a descriptive behaviour because they store information about the costs, the date of the study and an additional description. It is possible to store a number of attributes that have been investigated in a certain study. A *study* is also associated to a certain *study type*, comparable to a biobank (*participantDB*). We also provide the possibility to derive certain cases of available studies, which enhances query capabilities in a great way.

Possible research questions that can be answered with our model are

- Find cases which have been studied in an obesity study and which investigated the attribute BMI.
- Find all databases which have stored an obesity study that has been issued between 1999 and 2005

### Supported information structures

The keynote was to hold the structure very flexible in order to incorporate several aspects of different kinds of biobanks. For example:

- Cases can be subject related
- Cases can be only sample related
- Storage of multiple measurements (e.g., diagnose, temperature curve...)
- Different availability of attributes
- Existence of attributes and actual values
- ...

To accomplish those aspects the idea is to split up information into the classes *subject*, *case*, *sample* and *measurement*.

Via the relation *stores*, a biobank can upload their pseudonym unique identifier for cases. Depending on the type of a biobank (population-based, disease-oriented pathology) uploading subject ids is relevant or not. Therefore, the class *subject* is an optional feature, that means one can upload a subject id to certain cases, but it is not a mandatory requirement. The identifier of subjects must be a pseudonym and only the local biobanks are able to map that pseudonym identifier to a certain person.

With the relation *sample-belongs-to-case*, it is possible to model several views of cases:

- a *case* can contain a set of *samples* of several types. For instance a case belongs to a set of samples which were collected over three months, or
- a *case* belongs to a one-time delivery of an amount of *samples* or
- a *case* may belong to no *sample*. For instance, the *case* is only *subject* related.

A *Measurement* is a record of jointly derived attributes, i.e. a set of attributes (with content and existence property) which belong together. With the relations *Measurement-containsExistenceAttribute-Attribute* and *Measurement-containsContentAttribute-Attribute* it is possible to simulate a look-up data mart as a kind of index in order to form the set of related attributes. The relations *sample-contains-measurement*, *case-contains-measurements* and *subject-contains-measurement* associate the appropriate set of related attributes to a specific case, sample or subject.

The dynamic declaration of attributes and the dynamic specification of measurements offer flexibility in the representation of a biobank. E.g., with this dynamic behaviour *biobank-A* can specify attribute *Gender* as a content-attribute while another *biobank-B* stores attribute *Gender* as an existence-attribute. The decision is up to each biobank as long as the attribute catalogue provides possible values for the chosen content-attribute.

The choice of content- and existence-attributes could affect requests on a certain material:

- **Requests on existence of attributes:**  
For a request on the existence of several attributes, it does not matter whether the requested attributes are declared as content-attribute or existence-attribute. The only fact to get a query-hit for that request is that the searched attributes are declared by a biobank.
- **Requests on content of attributes:**  
For a request on the content of several attributes, all requested attributes must be declared as content-attribute by a biobank in order to get a query-hit. E.g., Figure 4 shows an exemplary declaration of content- and existence attributes for a BBMRI hub within the network. A request on *male* patients who suffer from *C50.8* would not get a query-hit from *BB-y* because the attribute *PatientGender* is only declared as existence-attribute and thus has no information about its content.

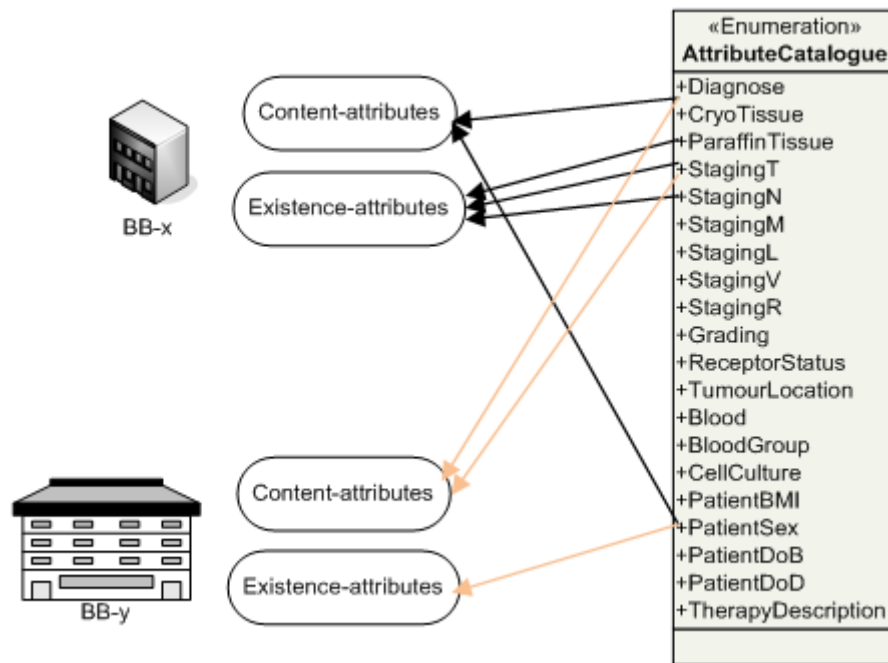


Figure 4: Explicit declaration of content- and existence-attributes

### 1.3.1.3 User, Role and Access Management

This part of the data model is used to control that the access of data on a regional hub. It allows managing users, roles and operations. Only the responsible administrator of a hub is able to create new users with an initial password. After the creation, the user gets roles assigned in the context of an arbitrary number of databases. Since it is possible that one participant is running several databases, we also have to ensure that a user may play different roles on those databases. E.g., Graz is hosting two databases where on one hand user-A is *administrator* of database-A, but on the other a *normal user* in database-B. To ensure the fulfilment of that needs each user gets roles assigned which always belong to a specific participant database. Connected to a role there are operations that the user is allowed to perform when owning a role. E.g., having the role *LocalAdmin* the user is allowed to add content to the participant database in which context he holds the role.



## 1.3.2 Prototype A – An early version of the generalized metadata model

### 1.3.2.1 Aim of Prototype A

The Prototype A of WP5 was realized as an explorative piece of software. The aim was to *validate requirements* (use cases and the derived workflow presented in Section 1.1.2) and to *check the feasibility* of the proposed architecture.

Almost as important as the aims of the prototype we have to state what the prototype was not planned to be:

- a fully functional software for a regional host
- a first version of a host, which only must be adapted to have a product ready.

### 1.3.2.2 Architecture

Within this section we discuss three different possibilities for solving the federation of BBMRI-Hosts. We end this discussion with an outline of the implemented functionality.

Within BBMRI, several regional hubs should act as research-platform to exchange knowledge. Within BBMRI, a hub can either be (1) an autonomous server on which surrounding biobanks can register or it can be (2) a biobank, which provides suitable interfaces and functionality. In (1) biobanks put the burden of implementing hub specific interfaces to the BBMRI hub. Biobanks only upload their contact and model information to the BBMRI regional hub. This helps especially small biobanks that are not permanently online in joining the network. In (2) big biobanks get the possibility to act as a BBMRI hub. This helps especially when bringing very large biobanks to the European network. Those biobanks usually have their own IT-infrastructure and get the possibility to join the network as a hub by implementing the needed interfaces. With these two different alternatives of regional hubs it is possible to bring a large number of biobanks into BBMRI. For the communication structure within the network of biobanks, three different approaches have been investigated. In the following, we provide a short summary with pros and cons of each approach.

- *Peer to Peer.* Within this approach, all participating biobanks act as hubs and are connected via a peer-to-peer infrastructure. All biobanks must provide a query interface because queries are sent to all participating biobanks by the requestor. The exchange format can be defined or even undefined. An undefined exchange format leads to interoperability problems - but defining an exchange format upfront could lead to a domination of the biobank with the smallest schema. Therefore defining the format is a critical success factor for the peer-to-peer architecture. One disadvantage of this approach is that small biobanks, which are not permanently online, cannot take part in the federation. The major drawback of the peer-to-peer approach is that there exists data in biobanks that is not allowed to leave the biobank until it is aggregated and anonymized.
- *Centralized with Integration Hub.* The centralized integration hub works as a mediator within the federation. This central hub distributes queries within the federation and it is also responsible for integrating results from the different biobanks. This architecture is also able to solve the disadvantage of the peer-to-peer approach with the use of a

data warehouse. Here the data is stored in an aggregated and anonymized form and it is available for the federation. Further on small biobanks could export the needed data and thus do not need to be online all the time. Nevertheless the centralized integration hub represents a single point of failure.

- *Combined Approach.* To overcome the before mentioned problems we designed an architecture for the collaboration between different biobanks as a hybrid of peer to peer and a hub and spoke structure. Within this approach, a regional hub (autonomous server or biobank) uses a meta structure to provide data sharing. All regional hubs (and participating biobanks that act as hub) are connected via a peer-to-peer structure and communicate with each other via standardized and shared interfaces. Participating European biobanks (without hub function) are connected with its specific regional hub via hub and spoke-structure.

Based on the considerations above the decision was made to follow the combined approach within BBMRI. Figure 5 shows a possible architecture for the combined approach. The shown architecture was also the starting point for the implementation of Prototype A.

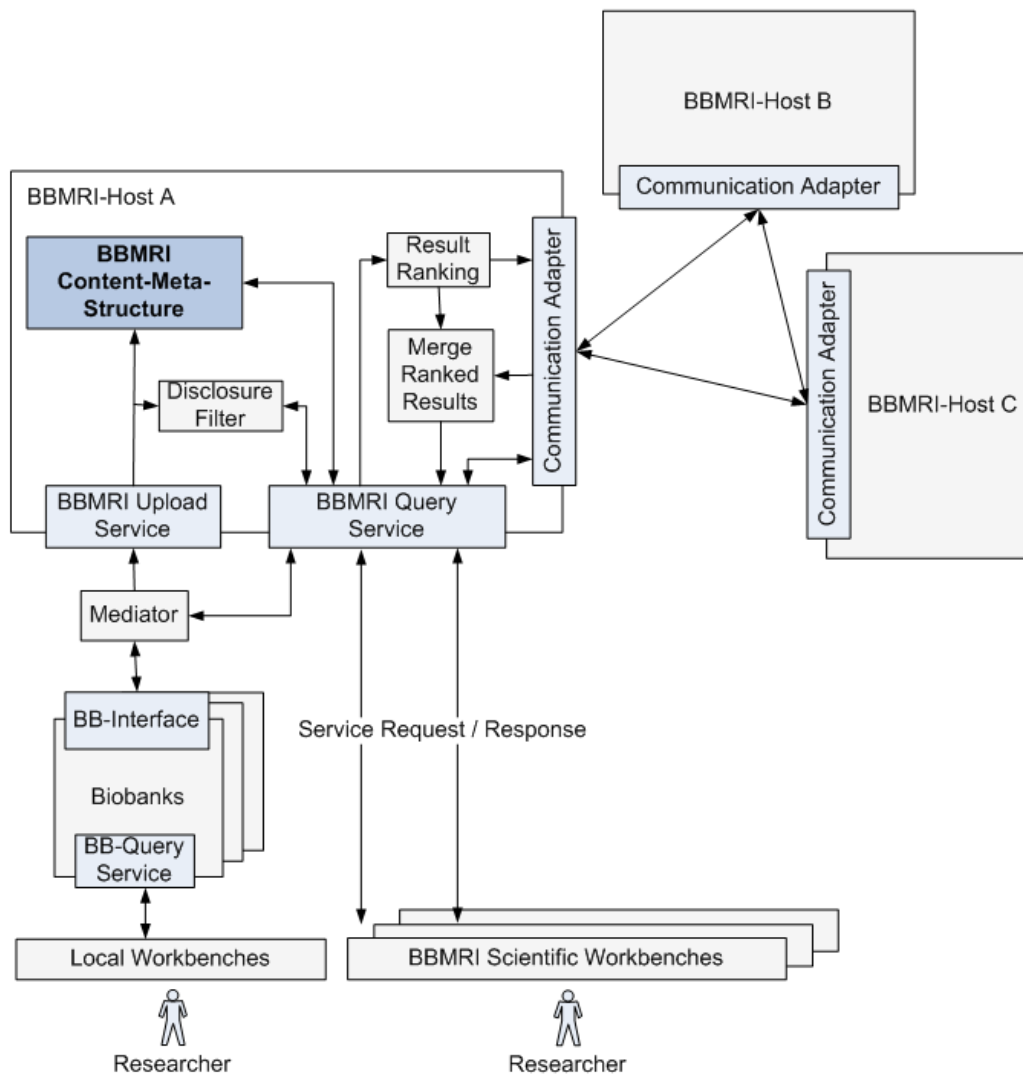


Figure 5: Proposed architecture for the combined approach

Within the prototype the functionality of *only one* BBMRI-Host (see Figure 5) was implemented, thus no communication support between different BBMRI-Hosts is present. During implementation of the prototype, the focus was on providing upload- and query-functionality. The aim was to check the feasibility of our scenarios and use cases as well as validation of the meta-schema presented in Figure 6.

Compliant to the concepts developed by WP5 and outlined here, a web application for WP3 (BBMRI catalogue) has been created, which is providing a broad overview of some of the European biobanks. This catalogue application comprises three layers: A service layer handles connections to repositories for data persistence and provides access to services for additional data processing functionalities. In a process layer, the interactions between the services are orchestrated. On the application layer, portal applications are accessible via a web interface (called scientific workbenches in Figure 5).

To query and visualize data from different biobanks, the WP3 Catalogue application was extended by an interface to the WP5 Demo Prototype. The user can determine search criteria by selecting values for attributes that are available at the BBMRI-Host. The specified query is forwarded to the service layer that handles authentication and identity issues and then sends the query to the BBMRI-Host. Via its QueryInterface the BBMRI-Host can be queried for appropriate database items of local biobanks, which stored their data in the metadatabase of the host. Finally, the query result is sent back to the web application where it is visualized for the user.

### **1.3.2.3 Some key facts and lessons learned**

Prototype A was implemented as collection of web services in java. The data exchange was realized via XML structure and SOAP (Simple Object Access Protocol) request. For simplicity, the metadata model (see Figure 6) was implemented in a MySQL DB.

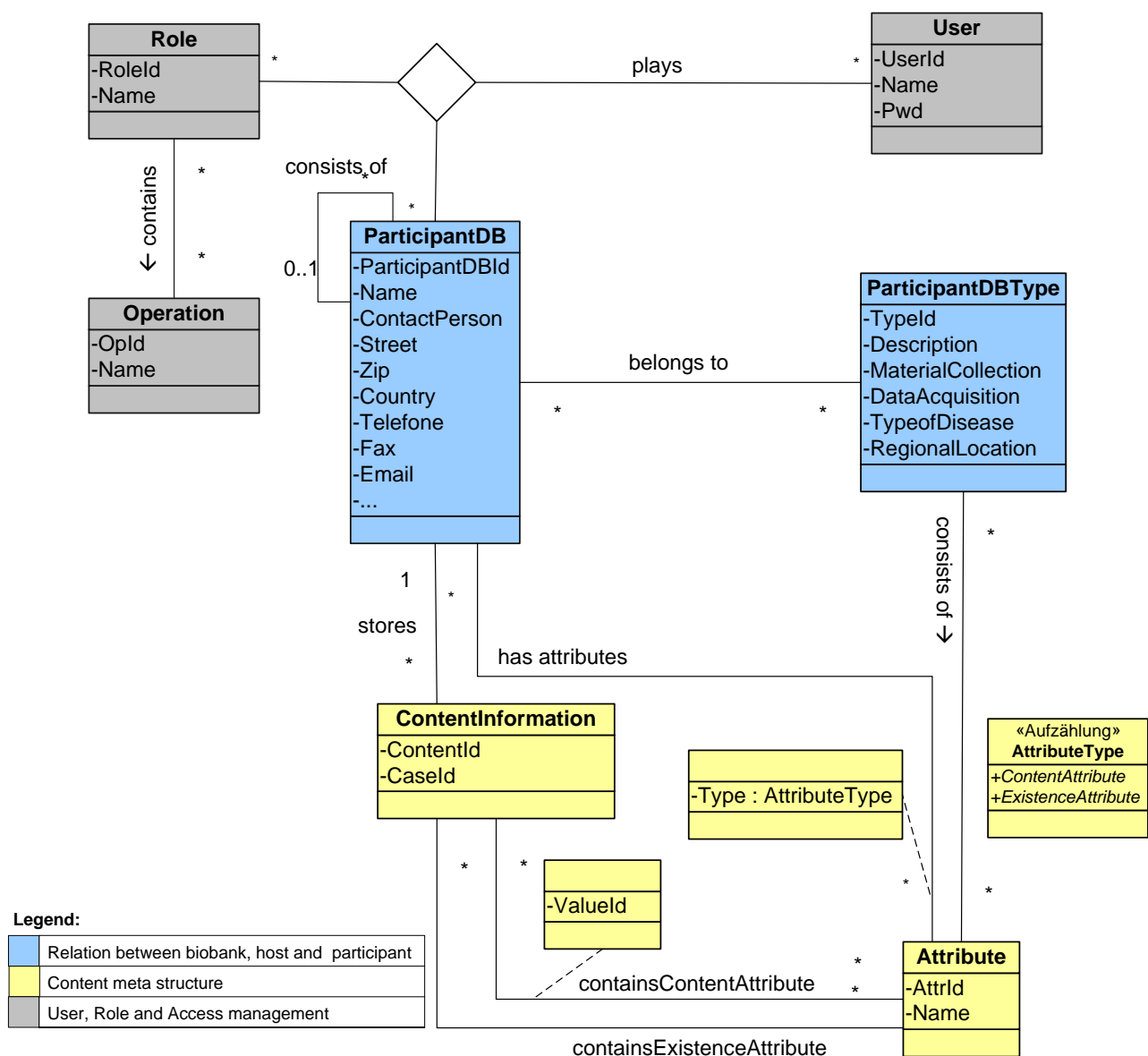


Figure 6: Implemented metadata model for the BBMRI-Demo prototype

With the first almost stable draft of the prototype, we started a testing phase and invited interested biobanks to participate and register for the prototype.

During this phase *eight accounts* have been released:

- 4 biobanks uploaded data (TU-Munich, MUG-Biobank, Institute for Molecular Medicine Finland, CRIP)
  - ~15.000 cases already stored in database
- The other account holders only had query-access-rights (Vitrosoft, CRB-IST, Karolinska Institutet Biobank, French National Cancer Institute)

We used this testing phase as a kind of feasibility analysis. In the following we present some lessons learned from the prototype:

(LL-1) The participation to the prototype was almost possible with little effort for the biobanks.

(LL-2) All in all it was inserted over 1 million of cases in a local environment.

(LL-3) Queries against this database were promising and had short response-times.

(LL-4) Thus we learned, that the introduction of use case 1a and 1b promises solid support in querying information provided by biobanks in the metadatabase.

(LL-5) Based to this experience we can suppose that the same applies for use case 1c and use case 2a and 2b without current implementation.

(LL-6) The usage of the prototype showed us important issues in case of

- representation and association of subjects
- in cooperation of different aspects of biobanks
- structure of obtainable attributes in the metadata

→ See Section 1.3.1 for enhancements of the meta-schema on which the prototype was built on.

(LL-7) The prototype was a good tool to check the requirements and how they are fulfilled.

(LL-8) It shows the feasibility of participation for a small set of biobanks.

(LL-9) The resulting rationale is that there is much more work to do when incorporating all the lessons learned into a real software infrastructure for BBMRI.

### 1.3.3 Prototype B – Based on the Set Definition Language (SDL)

Prototype B was developed in parallel with Prototype A. The related schema and proposed architecture have been presented as part of D5.3 and the documents cited there within. However, since neither the schema itself nor the architecture constitutes the major features within the proposal, only the prototype is presented below.

This prototype should not be considered as an attempt for a final implementation, although, in reality it is not far from a fully functional system that could be deployed globally. Rather, it was created to give a good understanding of the benefits of the SDL [13] design philosophy where the system is based on a flexible ad-hoc query language.

The main benefits of this prototype are that it is more or less data schema independent. The schema only needs to adhere to the SDL object relation model. This prototype provides the users with maximum query expressivity and the analysis is not restricted to a finite set of pre-designed queries. Likewise, it is easy to extend the schema, add new data types, or introduce new ontology without having to modify the system.

This prototype allows two types of queries, i.e., distributed metadata and subject aggregation queries. Detailed reports with data for individual subjects are only possible on a local node where the user has full privileges. These queries are envisioned as a mechanism to help the user to define the queries and reports that are eventually evaluated in a federated manner as subject aggregate queries.

This prototype simulates a federated query mechanism on the client side. It is however a straight forward task to move this code to the server side and collect the biobank reports by aggregating results from multiple servers for various network topologies.

It is assumed that each research user has access to BBMRI through a local node. In this local node, the user has full access to subject data, i.e. he is capable to issue queries for use case 4 in Section 1.1.2. If this would not be considered acceptable, a fake node with example data (like the one used in this prototype) could always be setup for this purpose.

The user accesses the BBMRI service through a web page that contains an SDL applet. Each server node must run an SDL server and a relational database. The web page with the prototype is shown in Figure 7. Currently no user authentication is implemented in the system. This would be relatively easy to add using

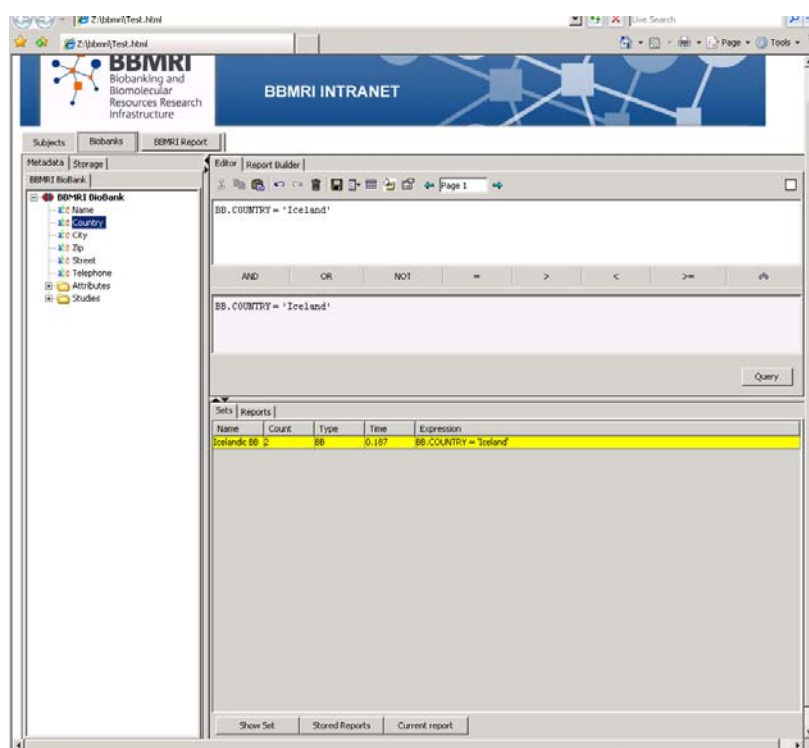


Figure 7: Web page for the SDL prototype.

browser based authentication standards such as OpenID [2] and the login does not have to be an integral part of the SDL applet although that is also feasible.

The applet has two standard SDL query windows, organized under a Biobank tab and a Subject tab, and a BBMRI report button to generate federated report. The usage patterns are then as follows:

- Define sets of biobanks and reports based on the BBMRI biobank data, i.e., the publicly available data on the biobanks that has also been referred to as the biobank meta-attributes (named BB.\*). The biobank sets can later be used to determine the scope of the federated queries.
- Define sets and reports based on detailed data from subjects (or samples). The user can evaluate the sets and the reports on his local node to shape the queries and get a “feeling” for the outcome. The sets and the report definitions are then used as the input to the federated queries. Thus, it is assumed that each BBMRI node supports a common minimal set of subject attributes (SU.\*), i.e., implements a common subject reference schema.
- Evaluate federated BBMRI reports. The user can decide to evaluate the query against all biobank nodes or only the biobanks defined in a biobank set selected from the Biobank query log. The user can then select one or more subject sets from the query log and have those set definitions evaluated against each node, i.e., estimate the set-size in each of the nodes. Additionally, the user can get histograms for any subject attribute/property/measure that he may have defined in one or more subject reports. Before the histograms are presented, *k*-anonymization is processed dynamically for the report that the user requests. Since users are typically only investigating a few attributes at the time, such *k*-anonymization does not have to perturb the data as much as when datasets with multiple columns (attributes) are released. Likewise, the dynamic *k*-anonymization is very fast.

A few screenshots explain this best. They can be found at: [www.decodevideo.com/bbmri](http://www.decodevideo.com/bbmri) in particular the link [http://www.decodevideo.com/hakon/bbmri\\_ex5.htm](http://www.decodevideo.com/hakon/bbmri_ex5.htm)

### 1.3.3.1 BMI analysis example

The screenshot in Figure 8 shows a report definition a user has setup in order to evaluate the maximum BMI value for each subject. This report has been evaluated for three different subject sets: all, males, and females.

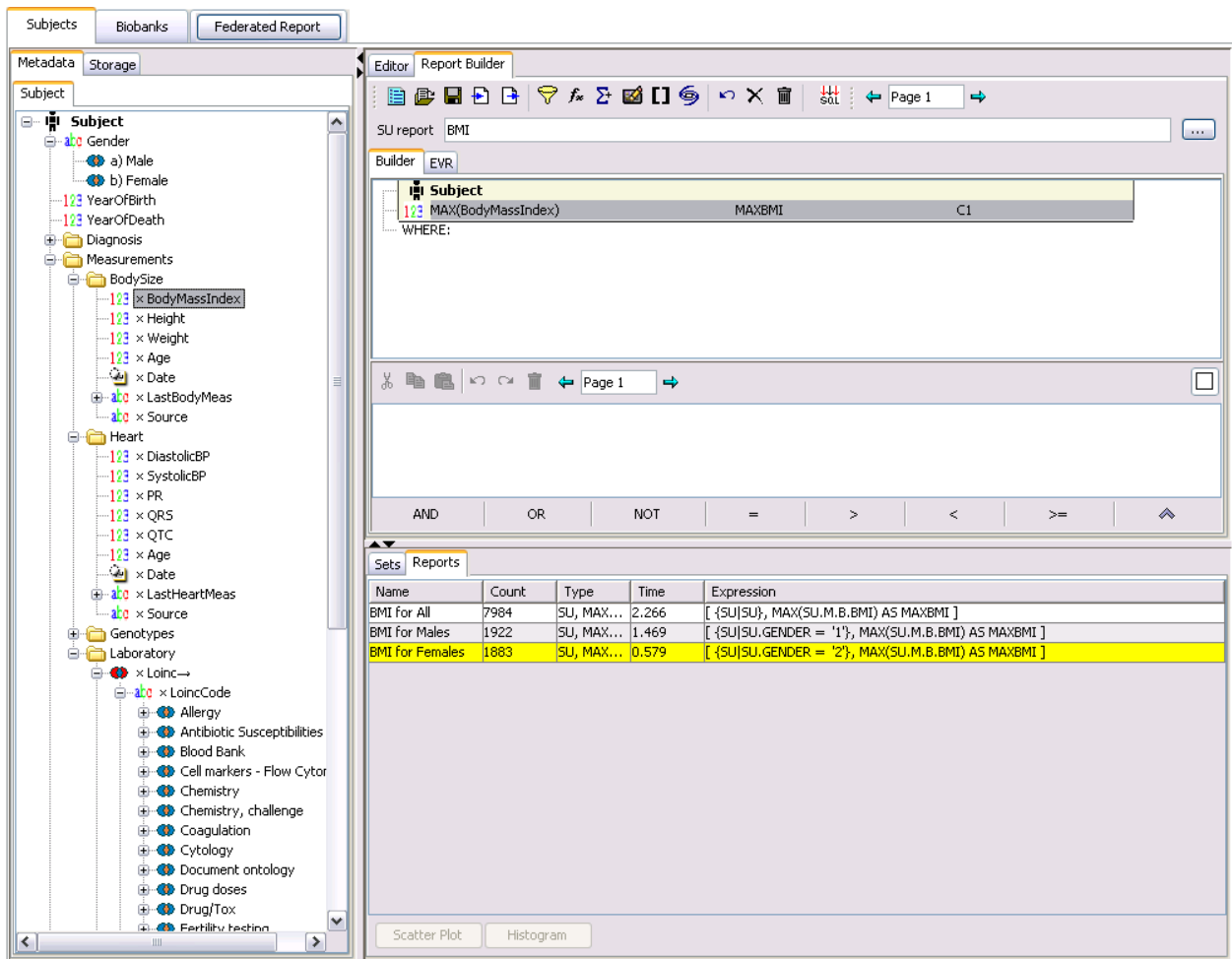


Figure 8: Screenshot from Report Builder.

Once the user wants to generate federated reports for all the BBMRI biobanks, he is shown the dialog in Figure 9 to select from sets and reports that he has specified in the session, sets and reports that he wants to evaluate remotely. Notice also the setting options for the inference algorithms (bin size and skew factor). In a real setup, these should not be accessible to the user but rather specified in each server that answers the queries.

Once the user has made his choice and pressed OK, the specifications are sent to all the BBMRI nodes. The servers return only skewed set-sizes and histogram bins and counts based on a dynamic  $k$ -anonymization algorithm. It should be pointed out that this inference algorithm uses deterministic skewing approach that prevents all tracker attacks.

The report window in Figure 10 shows an example of subject counts for the three sets and histograms for MAXBMI in three different fictitious biobanks, BB2, BB3 and DCGN. In this example the bin size and the skew factor were set to 5 and 2, respectively.

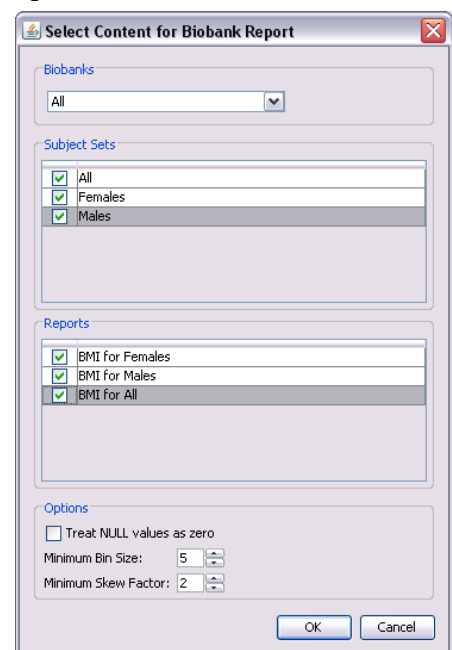


Figure 9: Content selection for reports.



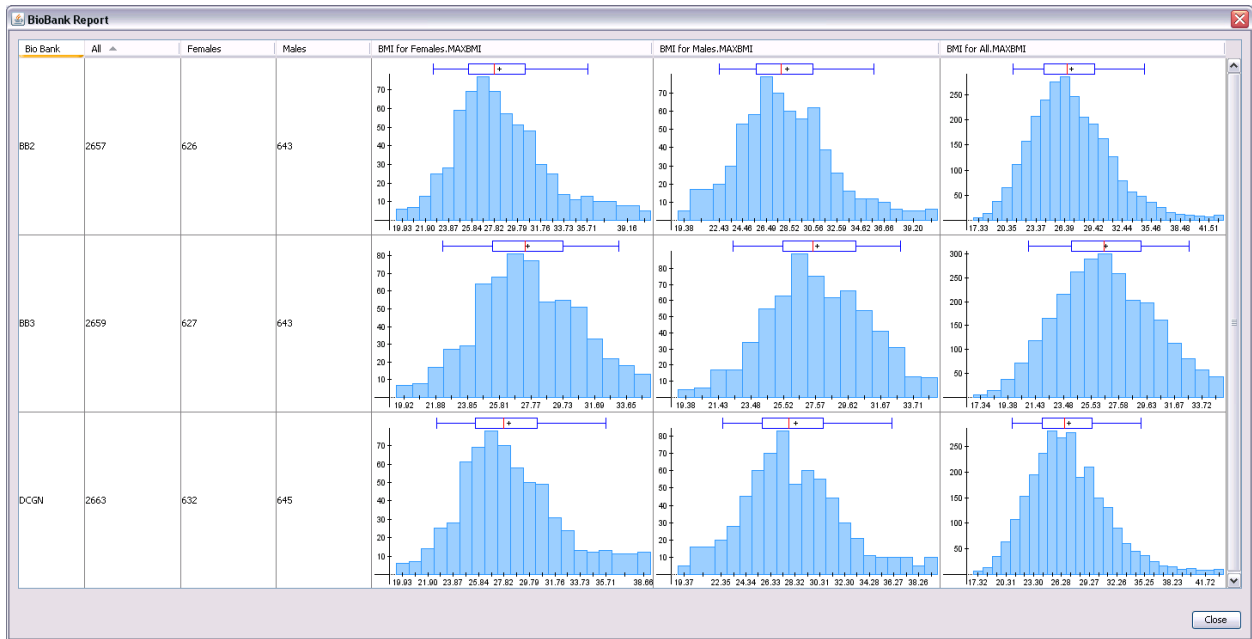


Figure 10: Example of subject counts for the three sets and histograms for MAXBMI in three different fictitious biobanks, BB2, BB3 and DCGN, with bin size = 5 and skew factor = 2.

The screenshot in Figure 11 shows the same federated report generated for the same specification as in Figure 10 but with a bin size of 15 and skew factor of 20. These numbers are unrealistically large and used for demonstration purposes only. Notice how the set-counts are differently perturbed. Also notice how one notices difference in the histograms clearly for the male and female sets (because they are based on fewer subjects) between the runs, whereas, for the combined set (All) the difference is less noticeable. The bars show the mean, median and the standard deviations of the distributions. They are hardly affected by the perturbation caused by the data inference protection algorithms.

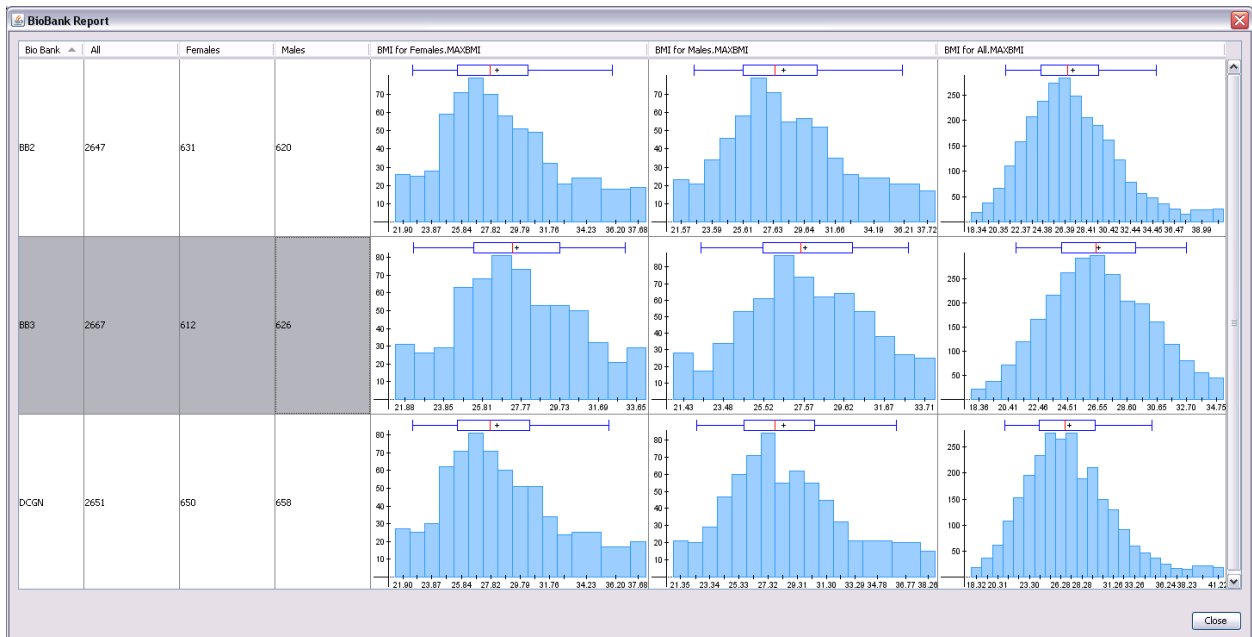


Figure 11: Same as Figure 10, but with bin size = 15 and skew factor = 20.

The last screenshot in Figure 12 simply demonstrates how users can dynamically specify new reports, how they can for instance investigate the difference in the distributions of the mean BMI and the maximum BMI of subjects.

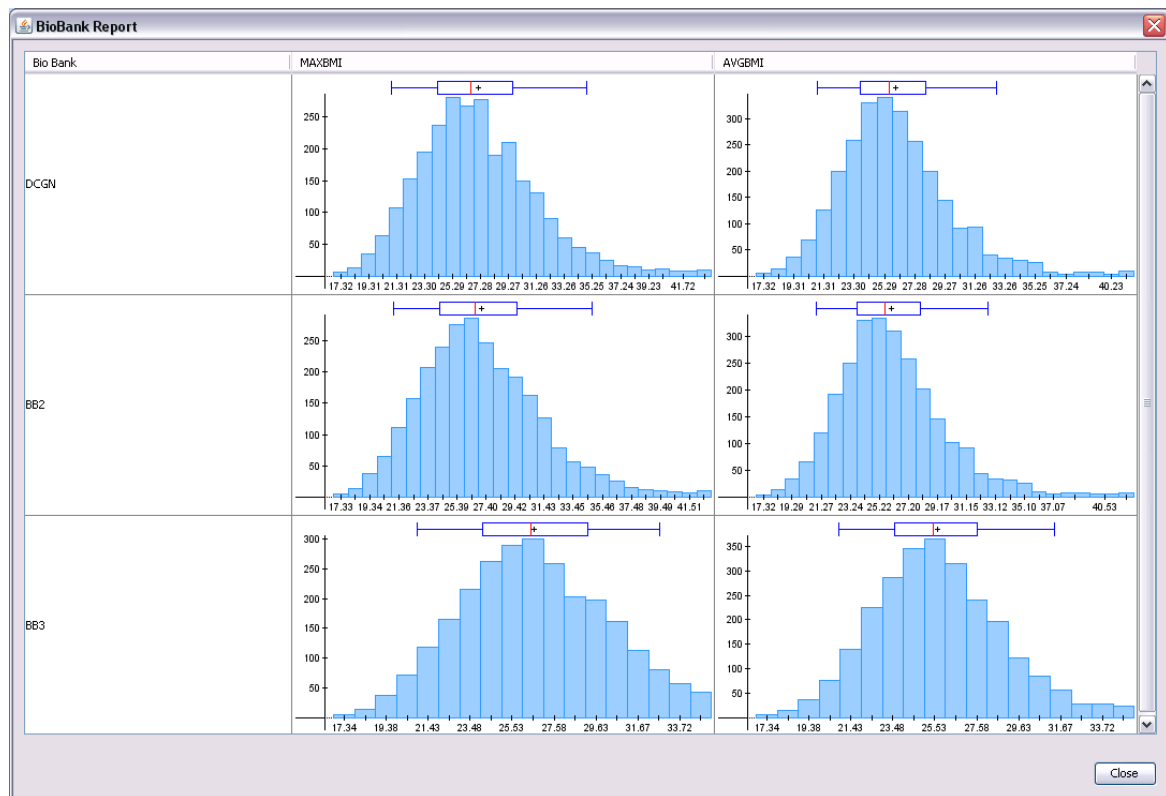


Figure 12: Showing the difference in distributions for the fictitious biobanks.

### 1.3.4 Scenarios for service architecture

The choice of an identifier (Section 1.2.1) needs to be evaluated in the context of the localization for data aggregation, and relates hence to issue of the service architecture. In D5.2 four different scenarios (Figure 13 and Figure 14) for the service architecture have been proposed and evaluated in relation to the use cases in Section 1.1.2.

Use case 1a and 1b in Section 1.1.2 can be adequately realized by architectures like shown in scenarios A and B (Figure 13). Scenario B architecture is comparable to A regarding autonomy and heterogeneity, but it handles distribution differently. In both scenarios, only metadata are stored in the hubs, and research queries are executed against the hub data. The hubs need to upload data from the local biobanks, which will result in a high number of local queries.

For use case 2, local biobanks need to be queried, and the local biobanks return pseudonymized identifiers. Scenario B can be modified as to return identifiers as well as aggregate data of remotely executed queries. Scenarios C and D (Figure 14) could be used to implement a solution for use case 2 as well, but they are different from A and B concerning data handling and autonomy: They return sample data to build semantic aggregates at the integration layer (e.g., aggregating samples belonging to a specific subject) and not at the biobank layer. In these scenarios, local processing needs are reduced while privacy and security have to be ensured, e.g., by anonymization. Use case 2 could be seen as a simple form of Scenario C or D. For any aspects of semantic integration, e.g., using the UMLS Metathesaurus, architectures from scenario C or D would provide advantages over scenario B.

It is important to note that if sample and/or subject data are sent to the portal, completely anonymized identifiers can be used as long as they are left unique. The necessary step of complete de-identification could be realized by means of a downloadable tool, by a client-side feature of the upload function (Scenario A-like), or as a feature of the biobank service (Scenario C-D). If samples need to be identified for shipping, the possibility to re-identify pseudonymized sample IDs would provide a benefit. Encryption of identifiers is an additional option.

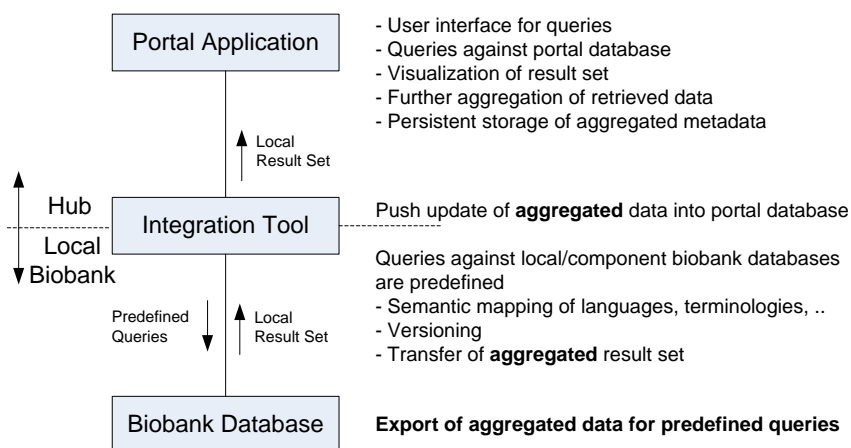
A main difference between scenarios A and B on one side, scenarios C and D on the other side is that in A/B local biobanks send aggregated data to the integration service while in C/D they send specimen data. Thus, C/D allow more sophisticated queries and require less local data processing, as A/B may result in high numbers of local queries. C/D requires additional security measures; this may include anonymization, which can be handled as mentioned above. In a WP5 phone conference in June 2009, there was a consensus that the current focus should be on approach B, whereas additional security measures would be needed for C, including both organizational and technical aspects. In the phone conference a disclosure model (also called ethical filter), *k*-anonymity and measures against trackers were addressed.

We have described two layers, abstracting from further hierarchical grouping. We have called the upper layer (level 2) the **hub layer** and suggested to run a **portal application** there. The concept can be directly used for the cooperation between national hubs (level 2) and local biobanks (level 1). In an approach with at least three layers, additional higher level hubs (level 3) would access these level 2 hubs, and it should be discussed that the level 3 hubs run a portal containing metadata (see scenario B). In this case of hierarchical layering, it is

recommended to add a **hub id** managed by the level 3 portal, completely analogously to the described architectural approach and scenarios.

In the WP5 phone conference in June 2009, there was also consensus that approach B appears to be the level of choice for higher-level hubs (at least for supra-national hubs), and that it can be expected that there will be biobanks who remain on level B, even if C has been introduced at a later point in time for the interaction between biobanks and lower level hubs.

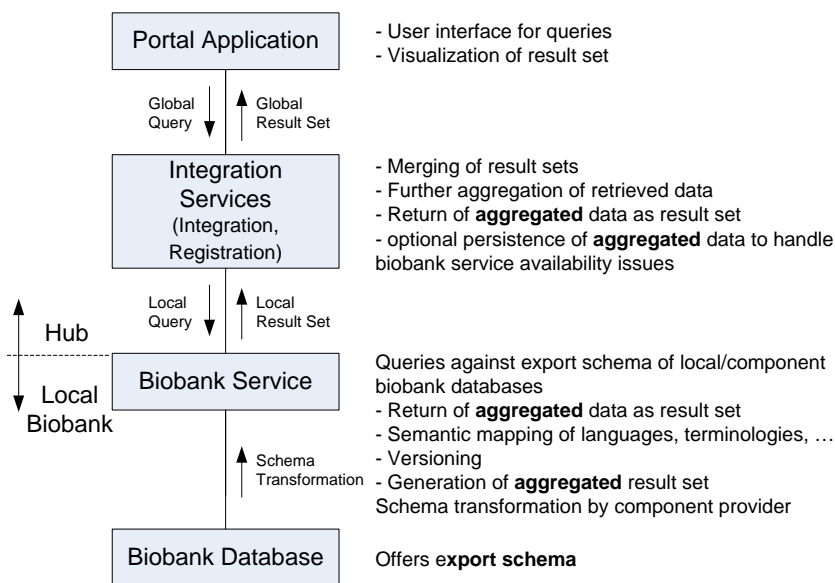
## Scenario A



- No external access to local biobank
- Update is performed when component biobank decides to send data
- If predefined queries are changed, all component systems need to modify their internal implementation
- Semantic changes (terminology...) demand changes of local implementations

Figure 1: Predefined queries, local aggregation of data

## Scenario B



- No external access to local biobank
- If agreed common attribute subset is changed, all component systems need to modify their internal implementation
- Semantic changes (terminology...) demand changes of local implementations

Figure 2: Federated schema, local aggregation of data

Figure 13: Scenario A (top) with predefined queries and local aggregation of data, and Scenario B (bottom) with a federated schema/shared data model and local aggregation of data.

## Scenario C

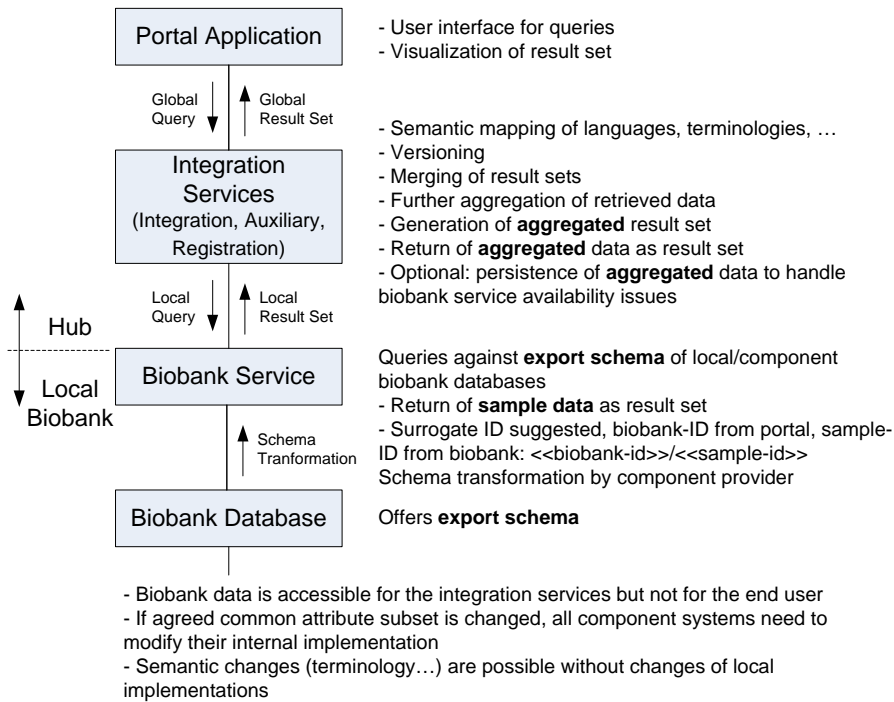


Figure 3: Federated schema, central aggregation of data

## Scenario D

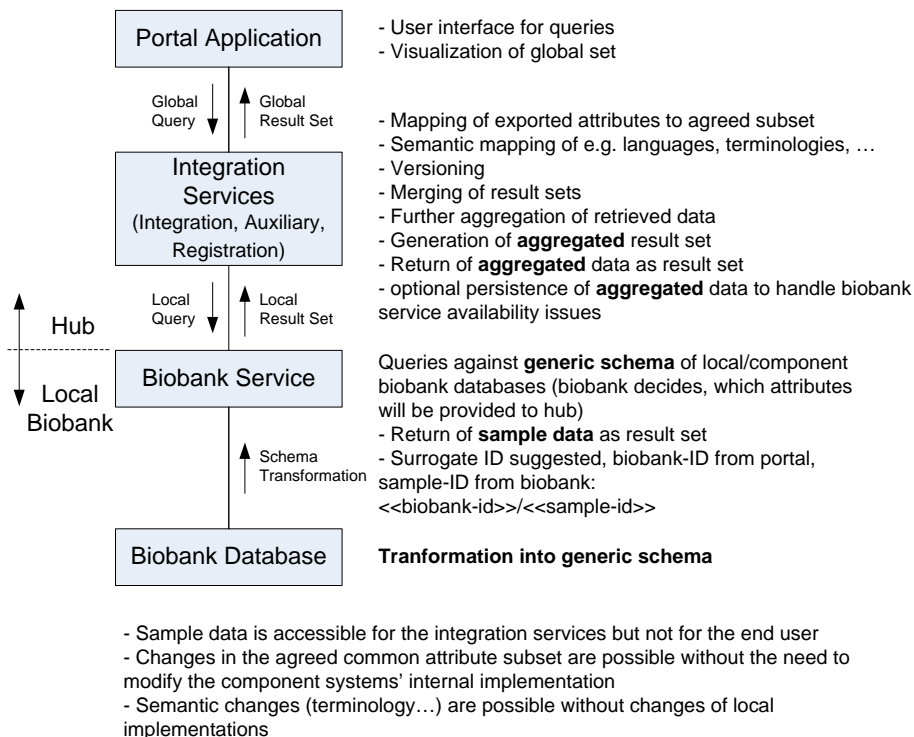


Figure 4: Generic schema, central aggregation of data

Figure 14: Scenario C (top) with a federated schema/shared data model and central aggregation of data, and Scenario D (bottom) with a generic schema and central aggregation of data.

### 1.3.5 Network and implementation model

A crucial point in systems design is the discovery and definition of key system entities. The definitions must be unambiguous and linked to the domain lexicon (R32). It is important to realize that the names of the entities do not necessarily have a one-to-one correspondence between lexicon and system models, because IT systems often have more stringent constraints. For example, the term ‘biobank’ may have slightly different meaning in system models than in the lexicon. An illustration of a possible network composition and implementation model is given in the next sections.

#### 1.3.5.1 Network model

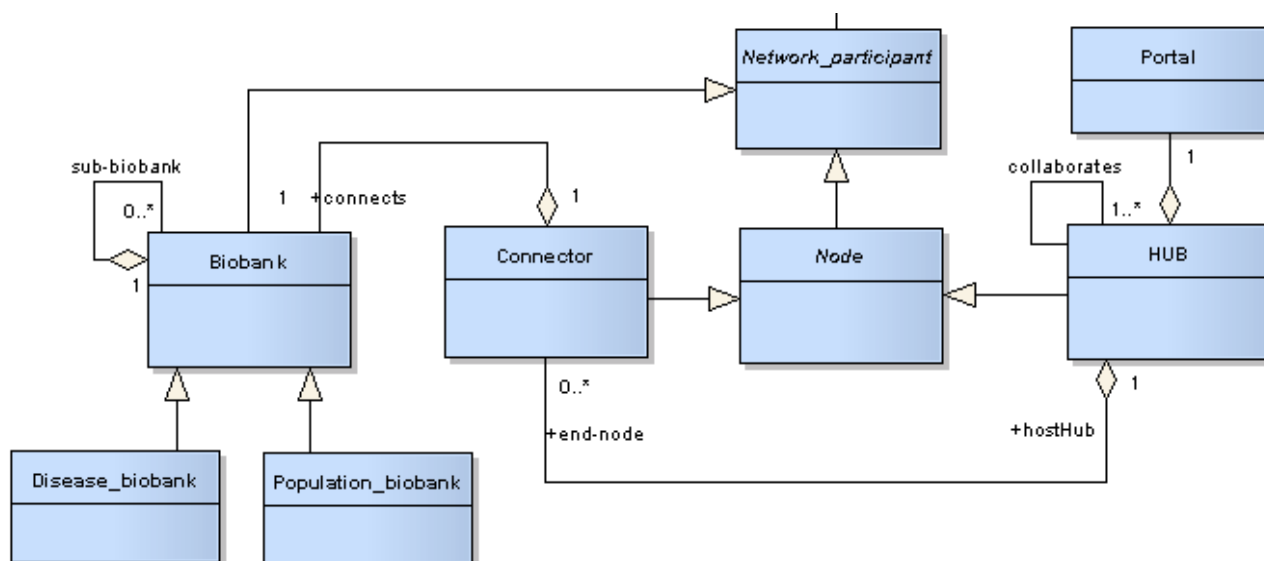


Figure 15: Conceptual network model – Network participants

A conceptual description of possible network participants is shown in Figure 15, using the UML modelling language. The Biobank network depicted has the following participants:

- *Biobank* (i.e., a BBMRI-Biobank) is an entity that owns and maintains samples and associated data. *Biobank* may have zero or many sub-biobanks collected for different purposes (*consist-of* relationship). There can be different kinds of biobanks, such as disease and population biobanks, which can have special properties of their own (specializations or *is-a* relationships represented by arrows in UML diagrams).
- *Connector* is a network node that provides API (corresponds to biobank service API) and connection services for biobank. Separation of biobank and connection services modularizes the network composition, and is a realization of the fact that some biobanks do not have enough resources for maintaining services of their own. *Connector* is defined here as serving only one biobank; i.e., it fully represents one and only one legal biobank entity, which clarifies the interface contract between the biobank and rest of the network. Different compositions are possible by internal arrangements; for example, external hosting services may host multiple *Connectors* as illustrated in Figure 16. *Connector* is connected to one hub only to make message routing simple. Rules can be adjusted later as the need arises. *Connector* can also participate in peer-to-peer connections assigned by a middleware service(see below), and can also provide the extended data federation services mentioned in D5.2. Commonalities between *Connector* and *Hub* should be defined and refactored into their own generalized classes. This option is illustrated in Figure 17.

- *Hub* is a network node linked to one or more *Connectors*. *Hub* also collaborates with at least one other *Hub*. *Hub* provides metadata index services, which can be used for resource discovery. *Hub* can also work as a mediator in data sharing scenarios (D5.2). Data sharing responsibilities can be adjusted later between *Connector* and *Hub* to address special requirements (R24). Local *Hub* should also take care of authorization and authentication because identity issues can likely be handled more reliably on local level (R10).
- *Portal* provides access to local and federated BBMRI data with necessary user management and authorization features. *Portal* can be implemented on *Hub* (D5.2) and/or portal applications can connect directly into the mediating, middleware (see below).

### 1.3.5.2 Implementation model - Federated Hub-and-Spokes network

The hub-and-spokes network model contains multiple end-nodes, connected by a hub which brokers between the nodes. The BBMRI network is composed of multiple national or local hubs (R18) connected together in a federated manner. Each hub connects one to many biobanks via a Connector module, as described in the previous section. The hub provides metadata query and data sharing services. Multiple hubs can be connected using a separate middleware application layer, which coordinates message passing between hubs (or nodes in general).

The network architecture is shown in Figure 15. A commonly used implementation pattern is the Enterprise Service Bus (ESB) model, which provides a message locator and queues for asynchronous service calls. In the model, service calls- which can be SOAP (Simple Object Access Protocol) messages- are not made directly to hubs (or other service providers), but sent into the bus instead, which delivers messages to specific targets and/or listener(s). Responses are sent back to the client application using the same message bus; responses can also be handles (URIs) to actual data). The bus can provide additional services such as message transformations and the registration services mentioned in 1.1.2 .

The interfaces of network participants should be defined and implemented for flexible composition and reuse as illustrated in Figure 16. For example: the Connector can function as an application module, which can be used by biobanks or hosting service providers. (The hub can also function as a hosting service provider without exposing this detail to the rest of the BBMRI network.) A goal of the modular design is to facilitate painless adaptation to the different implementation scenarios mentioned in Section 1.3.4. Applications should be open source to maximise the benefit of component-based development.



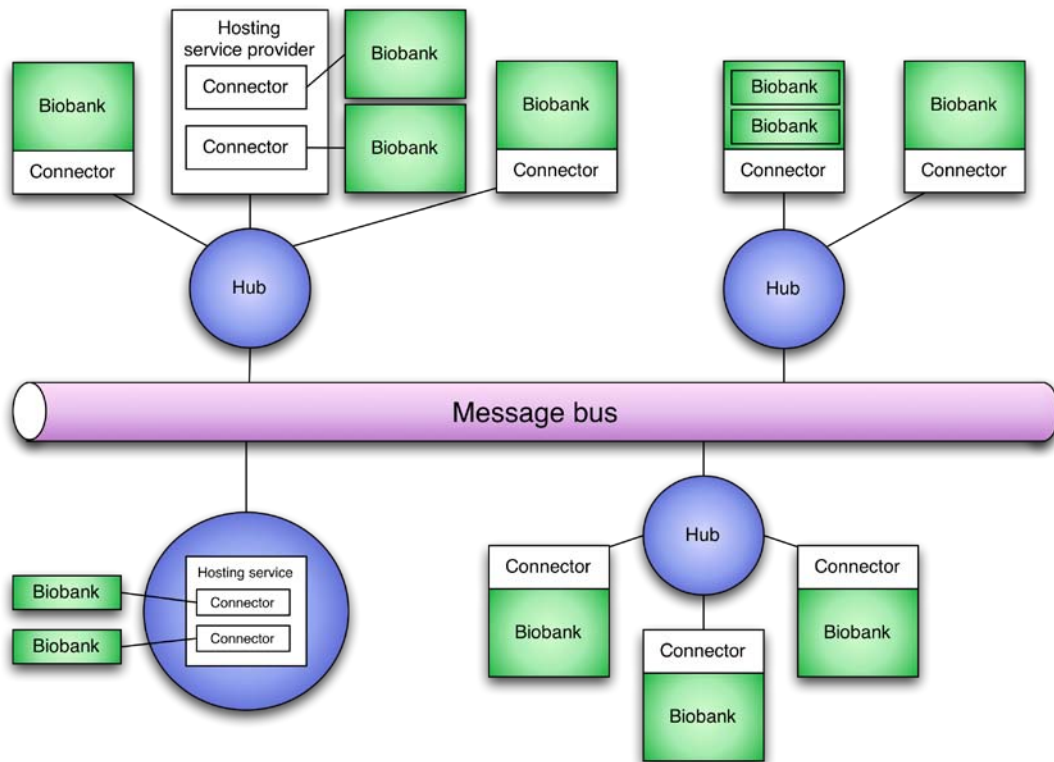


Figure 16: BBMRI Message bus. Hubs communicate with other nodes using a common message bus. Biobanks are connected to hubs via the Connector module, which can be part of a biobank, or can be in a separate hosting service. The hosting service can be external, or included in the hub's service package. Access to local and federated data is provided by portal applications connected to the message bus.

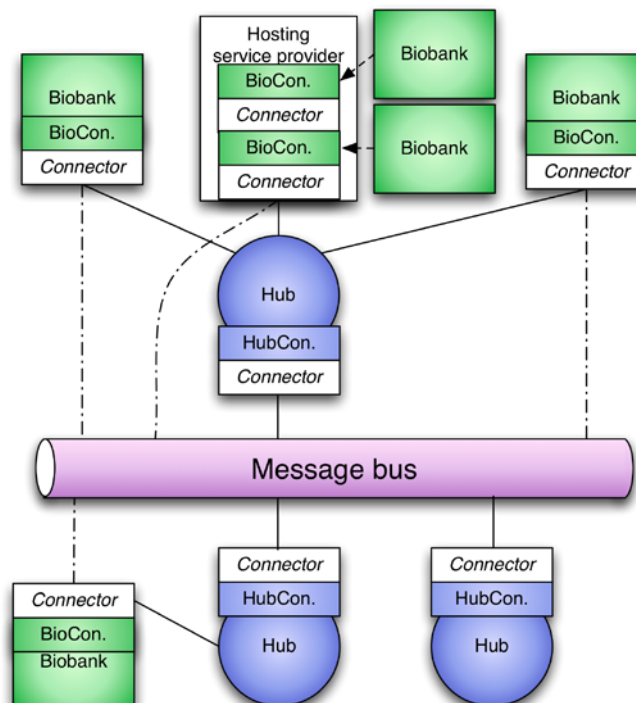


Figure 17: The Connector has been refactored into two specialized modules taking care of connection services for hubs (HubConnector) and biobanks (BioConnector). The Biobank connector module can provide services such as additional data federation capabilities directly from biobanks. The Connector module is generalized here to addresses networking needs, which are common to both hubs and biobanks.

### 1.3.6 The minimum data set

The minimum data set can be seen as an intermediate between the BBMRI questionnaires and the generalized metadata model presented in Section 1.3.1. The minimum data set is divided in to three levels; the biobank level, the study level and the object level - individual subject/case/sample. The idea is to provide an easy way to present which elements that are considered common in all biobanks. The minimum data set was originally designed at a WP5 meeting in Munich, December 14, 2009, but was heavily revised during discussions at the final WP5 meeting in Klagenfurt, February 4-5, 2010. It should be emphasized that it is a *minimum* data set, it has yet to be decided which optional definitions should be included. This must be done in collaboration with different domain experts.

#### *Data describing biobanks*

<b><u>Definition</u></b>	<b><u>Allowed values</u></b>	<b><u>Explanation</u></b>
BiobankAcronym	ASCII	
NameOfBiobank	Free text in English	
Institution	Free text in English	
URL		
Country	ISO-standard (3166 alpha2), two letter code	
ContactName	Free text in English	
ContactData	Free text in English	Address, Phone (E.164, No. 905 – 1.IV.2008), e.g., +46 8 524 877 59, Mail

#### *Data describing studies*

<b><u>Definition</u></b>	<b><u>Allowed values</u></b>	<b><u>Explanation</u></b>
NameOfStudy	Free text in any language	
EnglishStudyName	Free text in English	Translation of study name in English
ContactName	Free text in English	
ContactData	Free text in English	Address, Phone (E.164, No. 905 – 1.IV.2008), e.g., +46 8 524 877 59, Mail
KindOfStudy	Population-based, specific-disease, broad-spectrum of diseases	If "specific-disease", note ICD10
CategoriesOfDataCollected	[ClinicalDataAvailable, Diagnosis, Health information, Physiological/biochemical measures, Sociodemographic char., Socioeconomic char., Life habits/Behav., Physical environment]	Can be several values

*Data describing subjects/cases/samples within biobanks*

<b><u>Definition</u></b>	<b><u>Allowed values</u></b>	<b><u>Explanation</u></b>
AgeGroup	Interval [a,b], a>0, b<200, b>=a	a and b should be selected so that k-anonymity is guaranteed. Age group of donor at time for sample collection, number of age groups determined by biobank
Gender	Male, Female, Other	Gender of subject
SampleType	DNA, cDNA/RNA, whole blood, blood cells isolates, serum, plasma, fluids, tissues cryo, tissues paraffin-embedded, cell-lines	Type of sample. From the BBMRI core question.
SampleDate	ISO-standard (8601) time format	Date when sample was harvested
ClinicalDataAvailable	Yes/No	There exists clinical data related to the sample
OrganCategory	From the BBMRI Detailed descr bio samples	
OmicsDataAvailable	Yes/No	Genomics, proteomics etc
RestrictionsOnSampleUse	None, Consent participant, IRB approval, Approval of owner of collection	Can be several values

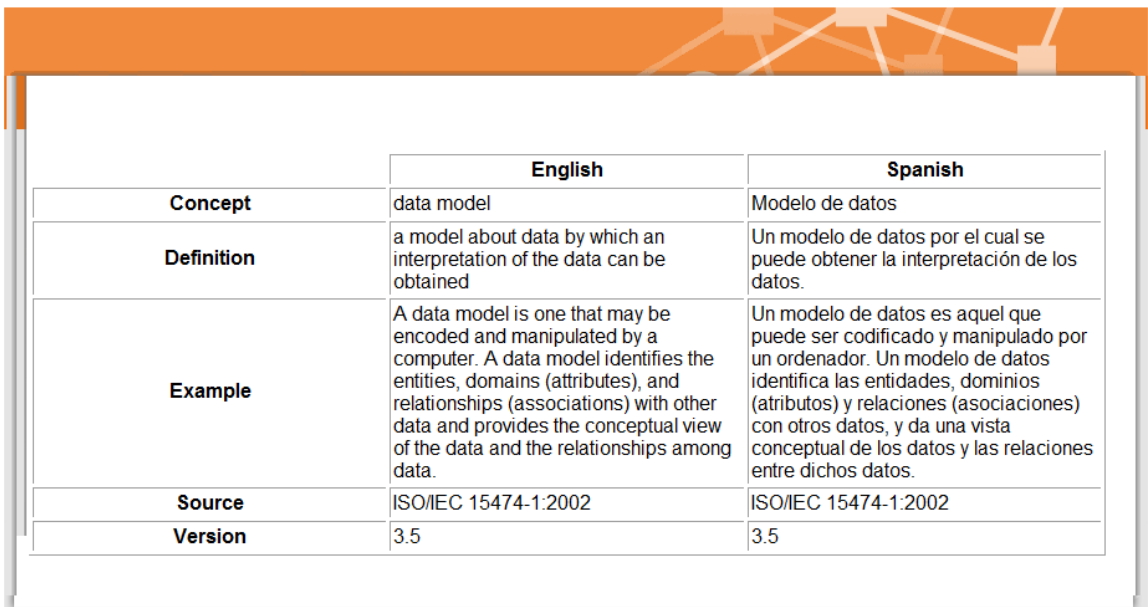
**NOTES:**

Time stamp and version control are part of the metadata schema and upload services

### 1.3.7 The Biobank Lexicon

The Biobank Lexicon is intended to unify interpretations of common terms in the domain of biobank informatics. At present, the vocabulary consists of 90 concepts with definitions. English is considered to be the master language with translations into six other languages; Estonian, Finnish, German, Italian, Spanish and Swedish.

An online version of the vocabulary has been implemented to facilitate use of the vocabulary, redefinition of concepts or additions of new concepts. The Biobank Lexicon Portal has been developed by using the open source Content Management System (CMS) called Joomla! (version 1.5.15) [14]. The system is based on PHP, CSS and JavaScript language and use the MySQL RDBMS system to store data. The middleware structure of this CMS allows the users to integrate different modules and plug-in. Moreover, the framework services ensure safety and simplicity for data management ensuring at the same time an advanced system for the portal security of the Biobank Lexicon [15, 16]. The Biobank Lexicon Portal home page is available at the URL: <http://www.biobank-lexicon.org> [16]. An example concept translation is shown in Figure 18.



Home > Translate Concept

## Translate Concept

	English	Spanish
<b>Concept</b>	data model	Modelo de datos
<b>Definition</b>	a model about data by which an interpretation of the data can be obtained	Un modelo de datos por el cual se puede obtener la interpretación de los datos.
<b>Example</b>	A data model is one that may be encoded and manipulated by a computer. A data model identifies the entities, domains (attributes), and relationships (associations) with other data and provides the conceptual view of the data and the relationships among data.	Un modelo de datos es aquel que puede ser codificado y manipulado por un ordenador. Un modelo de datos identifica las entidades, dominios (atributos) y relaciones (asociaciones) con otros datos, y da una vista conceptual de los datos y las relaciones entre dichos datos.
<b>Source</b>	ISO/IEC 15474-1:2002	ISO/IEC 15474-1:2002
<b>Version</b>	3.5	3.5

Figure 18: An example of concept translation at the Biobank Lexicon Portal.

## 1.4 Conclusions

### 1.4.1 Task 1: Requirements for a general information management system for biobanks in Europe

- Empirically derived *soft* requirements or *considerations* (Section 1.1.1):
  1. Data collection criteria could be used to estimate the quality level of information in databases intended for inclusion in BBMRI.
  2. The shared data model should be capable to deal with changes in data definitions over time, should make a separation of phenotype and genotype data and should have definitions of entities established in a multilingual domain lexicon.
- The following *use cases* have been identified, with decreasing priority and increasing complexity (Section 1.1.2):
  1. Search for biobanks.
  2. Search for cases.
  3. Statistical queries.
  4. Retrieval of detailed data.
  5. Upload or linking of data.
- Requirements have been formalized and categorised according to (Section 1.1.3):
  1. Technical requirements on BBMRI data integration system.
  2. Special data federation requirements.
  3. Networking requirements.
  4. Data schema and access requirements.

### 1.4.2 Task 2: Systems for maintaining unique and secure identities for specimens, subjects and biobanks

- Use cases 1 and 2 in Section 1.1.2 do not require globally unique identifiers issued and to be maintained by an external authority. Hence, surrogate identifiers, which should not contain any semantics, should be used. Exclusion of the semantic information from identifiers makes them more stable. It is important that identifiers can be created and managed locally in a coordinated fashion (Section 1.2.1).
- If need for a system for globally unique identifies should arise, ISO/HL7 OIDs will be a good choice as they are existing in the health care domain already. Mapping to the surrogates is possible by maintaining 1:1 mapping to surrogate keys, which are managed locally (Section 1.2.1).
- A final decision on a GUID standard for biological information should be made jointly with other affected ESFRI (Section 1.2.1).
- Researcher identification and user identification in general is another important issue. Emerging standards like Open Researcher and Contributor ID – ORCID (<http://www.orcid.org/>) should be used.

- Standard federated authorization and authentication protocols like SAML2 or OpenID must be used (R9).
- Work on the Data Protection deliverable is ongoing jointly with WP6.

#### 1.4.3 Task 3: Strategy for communication between biobanks, including a common nomenclature, compatible software techniques and appropriate information transmission policies

- *Data model and terminology*  
The generalized data model in Section 1.3.1 is a first step towards a data sharing in the BBMRI. The data model should be dynamic covering content and existence type of data. Expert group for the specific domain must define attributes based on standard vocabularies and multilingual Biobank Lexicon. Common set of attributes would define minimum dataset, which has been drafted in WP5 (1.3.6). Standard semantic web technologies must be used for building resource description frameworks for data and services. The data model is dynamic since each biobank may choose if a particular attribute should be of *content-* or *existence* type. The data model is adaptable to different kinds of biobanks by using different kind of schemas for *study types*. What attributes that should reside in a particular schema (e.g., for cancer biobanks) must be decided by an expert group for the specific domain. The common set of attributes for all study types would define the *minimum data set*, which has already been drafted in WP5 (Section 1.3.6). In order to obtain a unified view of the semantics of the attributes, each attribute, at least for the minimum data set, should be defined in a multi-lingual vocabulary, which would be an updated version of the *Biobank Lexicon* (Section 1.3.7).
- *Architecture and services*  
Architecture should be based on Service Oriented Architecture (SOA) pattern using standard data formats and application programming interfaces (APIs). Implementation should be based on standard web-service and grid technologies. The proposed architecture of database federation has been partially demonstrated by the two prototypes presented in Section 1.3.2 and Section 1.3.3. Prototype A also included a physical federation between biobanks in at least two different countries. It is the recommendation of BBMRI that the proposed federation architecture is kept as the major alternative for integrating data from various biobanks, to preserve the biobank autonomy. With this notion it has also been in agreement that for the near future it is most likely that only non-identifying meta- and aggregated data will be allowed to leave local databases, i.e., Scenario B in Section 1.3.4 would be the choice to implement at present.
- *Software technologies*  
The two prototypes developed are using different kind of software technologies. Prototype A is using XML-structured data for exchange by SOAP requests, with a data model implementation in a MySQL database. Prototype B is using the proprietary SDL and related software technology. BBMRI implementation should be made in a way that that different data access and query technologies can be used. The service APIs should expose data in a format that can be used by distributed query systems like SDL and convention web-service clients like in the web-service prototype.

- *Implementation strategy*
  - Standard web-service and grid technologies should be used leveraging existing frameworks like SDL, caGRID, I2B2 (<https://www.i2b2.org/>) and ACGT (<http://eu-acgt.org>).
  - Implementation should be based on open source principles, making sure that the framework ecosystem is based on same standardized components and/or reference implementations.
  - Application framework should be based on loosely coupled component model where different partners can easily contribute components for the common software ecosystem.
  - Architecture should be modular and implementable in a stepwise manner based on complexities and priorities. Recommendation is to layer the implementation based on type of data, which can be divided into public, non-sensitive metadata and individual level data. In this way, services can be built for non-sensitive data while working with complexities related to more sensitive individual level data.

## **1.5 External collaboration**

In order to harmonize other ESFRI engaged in the Life Science domain a meeting was held in Hinxton, U.K., November 16-17, 2009, between members of BBMRI and ELIXIR. On behalf of these two initiatives, a joint statement was drafted for future collaboration.

### **1.5.1 BBMRI/ELIXIR Working Group Statement: 16-17 Nov 2009**

**BBMRI Goal:** BBMRI's mission is to construct a pan-European biobanking infrastructure, building on existing infrastructure, resources and technologies, specifically complemented with innovative components and properly embedded into European ethical, legal and societal frameworks.

**ELIXIR Goal:** To construct and operate a sustainable infrastructure for biological information in Europe, to support life science research and its translation to medicine and the environment, the bio-industries and society.

There is a critical need to analyse and define the landscape and the communication channels between biobank resources, which will be federated under BBMRI and the integrated public data resources (such as the human genome in Ensembl), which are the responsibility of ELIXIR.

This challenge is large and incorporates standards, synonyms, data and process security; ELSI (Ethical Legal and Societal Issues) etc. To address this challenge and to make concrete steps going forward we have established a joint working group. The remit of this group will be:

- To understand and define the landscape for linkage, access and common querying between the biobanks and the public domain biomolecular resources (e.g., Ensembl Genome sequences)
- To analyse researchers' requirements for linking of resources to generate knowledge
- To encourage the development and adoption of common protocols from 'needle to freezer' and of a seamless provenance and quality management system

- To ensure that the same descriptions (metadata) are employed (or mappings provided) throughout for sample management
- To address the security issues which are relevant to the collaboration between BBMRI and ELIXIR, e.g., personal data and samples
- To promote and develop technical solutions to link between published results, ‘raw’ data and provenance
  - For access
  - For encouragement to publish/share raw data
- To provide guidance for the software solutions which will need to be developed to solve these problems, promoting open source solutions
- Establish a dialogue with stakeholders

Members of the BBMRI/ELIXIR Working Group:

BBMRI: Klaus Kuhn, Johann Eder, **Jan-Eric Litton**, Erik Bongcam-Rudloff, Martin Fransson, Eero Vuorio, Mike Taussig, Morris Swertz

ELIXIR: Paul Flicek, Alvis Brazma, Fiona Cunningham, Andrew Lyall, Nicola Slater, **Janet Thornton**, Ilkka Lappalainen



## 2 Deliverables and milestones tables

### 2.1 Deliverables (excluding the periodic and final reports)

TABLE 1. DELIVERABLES <sup>1</sup>									
Del. no.	Deliverable name	WP no.	Lead beneficiary	Nature	Dissemination level	Delivery date from Annex I (proj month)	Delivered Yes/No	Actual / Forecast delivery date	Comments
D5.1	Inventory of standard related issues.	5	5	Report	PU	12	Yes	13	Delayed one month. The approach with in-depth interviews was somewhat experimental. To benefit fully from the interviews it was decided that the deliverable had to be made more detailed than first planned.
D5.2	Strategy for unique and secure identities for specimens, subjects and biobanks.	5	5	Report	PP	18	Yes	18	
D5.3	Strategy for communication between biobanks including a common nomenclature, compatible software techniques and	5	5	Report	PP	20	Yes	25	Delayed five months since the deliverable need to (1) primarily be based on the outcome of D5.4; (2) include a report of the online Biobank Lexicon ( <a href="http://www.biobank-lexicon.org">www.biobank-lexicon.org</a> ), which was not

<sup>1</sup> For Security Projects the template for the deliverables list in Annex A1 has to be used.

	appropriate information transmission policies.								part of the original deliverable and (3) include comments on the WP5 prototype in the context of the WP3 Biobank Catalogue ( <a href="http://www.bbmriportal.eu">www.bbmriportal.eu</a> ).
D5.4	Requirements for a general information management system for European biobanks.	5	5	Report	PU	22	Yes	23	
D5.5	Strategy for a federated hub and spoke structure for European Biobanking.	5	5	Report	PP	24	Yes	25	
D5.7	Data Protection working group	5,6	5,6	Report	not specified	18	No	XX	Additional deliverable

## 2.2 Milestones

No milestones specified in Annex 1 with WP5 as lead beneficiary.

### 3 Acknowledgements

The following individuals are acknowledged as official WP5 participants in the BBMRI Grant Agreement:

<b>Official participant</b>	<b>Nationality</b>
Jan-Eric Litton (PI), Karolinska Institutet	SE
Antonio Fernandez, VITRO Ltd	ES
Hans Hillege, University Hospital Groningen	NL
Johann Eder, University of Klagenfurt	AT
Juha Muilu, National Institute for Health and Welfare	FI
Luciano Milanesi, Institute for Biomedical Technologies	IT
Paolo Romano, Istituto Nazionale per la Ricerca sul Cancro, Biological Bank and Cell Factory	IT
Paul Flicek, EMBL-EBI	GB
Tim Peakman, UK Biobank Ltd	GB
Tore Risch, Uppsala University	SE
Unnur Thorsteinsdottir, deCODE genetics	IS

In addition to the names above, several other people have contributed to the work presented in this final report:

<b>Participant</b>	<b>Nationality</b>
Christian Koncilia, University of Klagenfurt	AT
Claus Dabringer, University of Klagenfurt	AT
Michaela Schicho, University of Klagenfurt	AT
Christina Schröder, IBMT Fraunhofer	DE
Dominik Schmelcher, TU Munich	DE
Gregor Lamla, TU Munich	DE
Karsten Heidtke, IBMT Fraunhofer	DE
Klaus Kuhn, Technische Universität München	DE
Oliver Gros, IBMT Fraunhofer	DE
Sebastian Wurst, Technische Universität München	DE
Erkki Leego, University of Tartu	EE
Fernando López, VITRO Ltd	ES
Jonathan Horan, VITRO Ltd	ES
Pedro Roiz, VITRO Ltd	ES
Blandine Rimbault, Institut Pasteur	FR
Fanny Jadeau, Institut de Biologie et de Chimie des Protéines	FR
Ivan Kergourlay, CHU de Rouen	FR
Lise-Marie Daufresne, Institut Paoli Calmettes	FR
Louis Rechaussat, INSERM	FR
Andy Harris, UK Biobank Ltd	GB
Fiona Cunningham, EMBL-EBI	GB
Maria Krestyaninova, EMBL-EBI	GB
Mario Caccamo, EMBL-EBI	GB
Hákon Gudbjartsson, deCODE genetics	IS

Andrea Calabria, Institute for Biomedical Technologies	IT
Gerard Van der Hoorn, String-of-Pearls	NL
J. Jaap Nietfeld, University Medical Centre Utrecht	NL
Jan Talmon, String-of-Pearls	NL
Álvaro Martínez, Uppsala University	SE
Erik Lagercrantz, Swedish University of Agricultural Sciences	SE
Fredrik Lindén, epSOS	SE
Martin Fransson, Karolinska Institutet	SE
Roxana Merino Martinez, Karolinska Institutet	SE
Ruslan Fomkin, Uppsala University	SE

## 4 References

1. Sweeney, L., *k-anonymity: A model for protecting privacy*. International Journal of Uncertainty Fuzziness and Knowledge-Based Systems, 2002. **10**(5): p. 557-570.
2. *OpenID Foundation website*. [cited 2010 May 4]; Available from: <http://openid.net/>.
3. Eder, J., et al., *Information Systems for Federated Biobanks*. Transactions on Large-Scale Data- and Knowledge-Centered Systems I, 2009. **5740/2009**: p. 156-190.
4. *International Organization for Standardization*. [cited 2009 June 8]; Available from: <http://www.iso.org/iso/home.htm>.
5. *HL7 OID Registry*. [cited 2009 May 5]; Available from: <http://www.hl7.org/oid/index.cfm>.
6. Paskin, N. *Digital object identifier system*. [cited 2009 May 4]; Available from: <http://www.doi.org/overview/080625DOI-ELIS-Paskin.pdf>.
7. *OMG Life Sciences Identifiers Specification (LSID)*. [cited 2009 June 8]; Available from: <http://xml.coverpages.org/lxid.html>.
8. *GEN2PHEN*. Available from: <http://www.gen2phen.org/groups/researcher-identification>, <http://www.gen2phen.org/wiki/standards>.
9. Gottweis, H. *Publics and Biobanks*. in *Biobank Symposium Graz*. 2009.
10. Gottweis, H., *Public perceptions of biobanks: Emerging Themes From Austrian-Dutch Focus Group Research*. 2009.
11. Byun, J., et al., *Privacy-Preserving Incremental Data Dissemination*. ACM Journal of Computer Security, 2009.
12. Malin, B., *A computational model to protect patient data from location-based re-identification*. ELSEVIER Artificial Intelligence in Medicine, 2007.
13. *The SDL system - Ad-hoc querying and reporting with a Set Definition Language*. 2010, deCODE genetics.
14. *Joomla!* ; Available from: <http://www.joomla.org>.
15. *Italian BBMRI site*. Available from: <http://www.bbmri.it>.
16. *Biobank Lexicon Portal*. Available from: <http://www.biobank-lexicon.org>.