Table of contents

Introduction

The services in Data4lifesciences WP3 will optimize discovery of, and access to, Dutch UMC data and sample collections by linking (local) registries ranging from local studies up to (inter)national networks.

Background

Currently, most UMCs have adopted a FAIR[1] data policy and support the FAIR data principles. BBRMI-NL has developed and manages a national sample catalogue. This catalogue currently still focuses on samples and is lacking information on data collections. Also, the sample information needs to be updated. On a local level, several similar directories exist but many sample and data collections are not included. This makes data and samples insufficiently findable and accessible for others hindering re-use of data and samples.

True North

Based on the listed observations regarding findability and accessibility of samples and data, we arrived at the following mission, formulated as a 'true north' (what, who, why, when, how):

- Ability to find and request all data and samples in the Netherlands. This includes a wide range of data, (*e.g.* registry data, phenotypic data, omics data) and samples from NL and beyond;
- By every researcher and clinician in NL and beyond;
- To unravel disease aetiology and inform (precision) clinical decision making;
- Now and forevermore;
- In a hybrid centralised/federated, horizontally scalable environment with attention for donor privacy/ethical concerns.

Objectives

This WP has two main objectives:

1. the WP will create a network of local partners for the data/sample access. This network will:
    - share best practises to create awareness in the local UMC on data/sample sharing and access;
    - develop a common vision, strategy and approach on how to implement FAIR data/sample catalogues in local UMCs and thereby contribute to FAIR data on a national level.
2. the WP will create a network of web-based catalogue tools comprising:
    - **Directory**: "yellow pages" for discovery of (summary-level) information on data and sample collections, e.g. biobanks, registries, image collections, and cohorts.
    - **Data/sample locator**: a functionality to be able to count the existence of suitable samples based on (individual-level) queries, *e.g.* age, disease, material, to be able to evaluate whether enough samples are available with the right characteristics.
    - **Data/sample request workflow**: a workflow system to coordinate sample and/or data access requests to dramatically reduce the barriers to request and receive access to (privacy-sensitive) collections.
    - **Mapping tools & metadata catalogue**: facilitate pooling of data from multiple studies/collections (*i.e.* union of the data, not record linkage) by cataloguing data dictionaries and assist mapping of data elements.

All services will have tiered access controls such that data providers can decide if the information is public access, requires a registered user, or requires an approved data access request.

[1] FAIR: Findable, Accessible, Interoperable and Reusable, see Wilkinson *et al.* Scientific data 2016

Results from the preparatory phase

WP3 builds on work that has been done in the context of BBMRI-NL and the CTMM TraIT project, as well as the BBMRI ERIC, BioMedBridges and other EU projects. In BBMRI-NL and BBMRI-ERIC the landscape of catalogues has been explored and we defined several levels of catalogues to make a clear distinction between different types of catalogues. We recognize five different levels ranging from a directory of biobanks at level 1 to a donor/patient level catalogue that allows for linking of samples with other records at level 5. Since 2011 BBMRI-NL has been operating a catalogue of biobanks at level 1, which currently contains over 200 biobanks. Meanwhile a level 5 catalogue and a request broker have been realized for the CTMM TraIT project and a sample locator has been realized for the Stichting PALGA. Federation of biobank catalogues has been piloted within the context of BioMedbridges, but as of now the landscape of biobanks is still fragmented without a single entry point for researchers.

Description of work

The work will be organized in the following components:

Set-up a network of catalogue partners from local UMC's

In each UMC a contact will be identified and for data/sample access and catalogue. These partners will be connected and coordinated by the WP leader. Between these local partners, experiences and best practises regarding data/sample access will be shared and documented. Building from these experiences and best practices a common vision, strategy and approach will be developed to implement the FAIR data/sample catalogue locally and create a national FAIR catalogue.

Development of catalogue data structures

This task will develop, maintain, and version the common data structure that is used within the catalogue network. In recent years, a first version of the biobank catalogue (catalogue level 1) has been provisioned within BBMRI-NL. This version enables sharing of biobank and collection information. Based on user demand from the Data4lifesciences community, the catalogue will be broadened. For example, there are already requests to add summary information on imaging metadata (in relation to Euro-BioImaging), rare disease data (in relation to RD-Connect), molecular data sets (with cross reference to other databases such as the European Genomics Archive (EGA)) and data dictionaries (from the BioSHaRE and BioMedBridges projects).

Data/Sample locator

This task aims to maximize usability and visibility of the catalogue services and impact for the UMCs. While the currently existing catalogues (BBMRI-NL, LifeLines, PALGA, CTMM-TraIT) provide the basic search capabilities, the search and display should be enhanced to improve the added value and attract more sample/data sources and users. The task will investigate the possibilities and needs to improve usability and visibility of the catalogue.

Data Federation Protocol

This task will enable the technical connectivity between the nodes in the catalogue network, *e.g.* the internal catalogues in the UMCs and the central BBMRI-NL catalogue. In a pilot with BioMedBridges and BBMRI-ERIC, a lightweight protocol has been developed to share summary level information between the existing catalogues of the national nodes and networks at the one hand and the BBMRI-ERIC central directory at the other hand. This system is based on REST + RSQL, which can easily be implemented by local IT departments. Also, other networks have

shown interest to generalize this approach as a standard tool for data federation/syndication, *e.g.* RD-Connect. However, currently each change in the data structure requires programming and deployment. In order to make the catalogue network sustainable, the protocol will be re-developed so it can be automatically configured upon changes to the data structure without requiring additional software engineering investments. Where possible, technical interfaces and tools will be shared and aligned with other Data4lifesciences WPs.

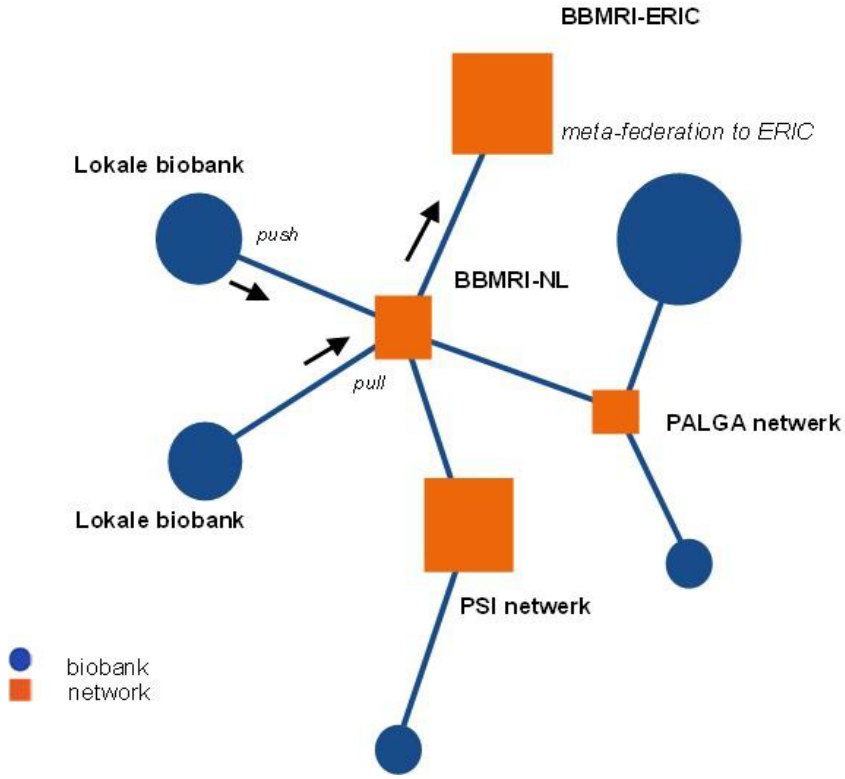

Figure 1. Schematic overview of local, national and international data/sample catalogues, using a federated model.

Connector Toolkit & Support

This task will ease the integration of new nodes into the directory and sample locator network. A pilot in BBMRI-ERIC demonstrated that nodes with IT resources can program the directory protocol into their software; however, not all UMCs have sufficient IT resources. Therefore, we will develop a 'connector toolkit'. The toolkit will consist of a software library that nodes (*i.e.* UMCs) can build into their own software and are ready-to-run. IT and data management support for this implementation will be required from the local UMC.

Mapping tools & Model registry

This task will integrate a data harmonization tool, for example using BiobankConnect (http://pubmed.org/25361575) developed between BioSHaRE and BioMedBridges, to map and transform data from local registrations into the BBMRI-NL data structure. Also, functionalities that are developed in the context of other projects, will be evaluated with NFU partners for implementation.

Request workflow for data and samples

Users having identified interesting data/samples in the catalogue can request those samples in a

sample request workflow for communication between researchers (consumers) and biobankers (producers) regarding requests. Such communication is expected to run in repeated cycles, with improving specificity of what the researcher needs. Also, the system will be designed to work across multiple nodes (so that one request can be partially fulfilled by multiple contributing biobanks). In the preparatory phase LifeLines has invested in a catalogue of data/sample items where interested researchers can explore these items, add a selection of items to a shopping cart, and subsequently submit a request to the LifeLines research office. In addition, a sample request workflow has been piloted for the PALGA pathology network in BBMRI-NL phase1.0.

Approach and Planning

Objectives 2017 - 2018

- The prime focus of WP3 is to set-up the partner network. With the network, a common vision, strategy, and approach will be developed for broad implementation of FAIR data/sample catalogues (access to data and samples) in the local UMCs.
- The IT focus of the WP is to deliver a national level 1 catalogue for UMC data and sample collections, based on the current BBMRI-NL level 1 catalogue.
- Where possible, the local catalogues will be linked to the national catalogues to enable automatic updates.*
- A request workflow will be connected to the level 1 catalogue, implemented for the local biobanks that are able to, and willing to, support such a workflow.*
- User interface improvements in order to improve the usage and usability: a more user-friendly 'look and feel' of the catalogue.
- Closely collaborate with each of the UMCs in order to further extend the data/sample content and improve the data quality.*
- Extend the catalogue with deeper information (subcohorts, data items, etc.) in close collaboration with the local data managers in the UMCs. Upon request, level 2, 3 or 4 catalogues will be offered to local data or sample collection owners.*

  * *This deliverable requires additional personal capacity at the local UMC partner, both data management staff and/or developers*

Deliverables 2017 - 2018

The work from this WP is done in collaboration with BBMRI-NL2.0 WP5.

Nr	Deliverable	Responsible	Planning	Capacity needed
1	Make an inventory of partners in UMCs and set up the partner network	Salome Scholtens	Q2 2017	UMCG effort
2	Make an inventory of local data and/or sample collections that could be connected automatically or manually to the national level 1 catalogue	Salome Scholtens	Q3 2017	Core team and local partners
3	Improve data quality of the catalogue; outreach to biobanks to update their own data in current catalogue	David van Enckevort	Q1-Q3 2017	Local partners
4	Add imaging and omics studies	David van Enckevort	Q3 2017	Local partners and Molgenis development team (UMCG)
5	Connect local data and sample connections (API and pipelines for automated updates).	David van Enckevort	Q3 2017	Local partners and Molgenis development team (UMCG)
6	Organise a user experience meeting in order to receive feedback on the usability and look and feel of the catalogue	Salome Scholtens	Q4 2017	Core team and local partners

Nr	Deliverable	Responsible	Planning	Capacity needed
7	Improve look and feel of the catalogue	David van Enckevort	Q4 2017	Molgenis development team (UMCG)
8	Integration with sample request workflow (Q4); proof of concept implementation with GoNL, BIOS, PSI, and local UMCs	David van Enckevort	Q4 2017	Molgenis development team (UMCG)
9	Document local experiences with data/samples access and best practices	Salome Scholtens	Q4 2017	Local partners
10	Develop a common vision, strategy and approach for local implementation of FAIR data/sample access	Salome Scholtens	Q2 2018	Local partners

Risks

1. Currently there is very limited budget for the local nodes to engage in connection activities.
2. Biobanks may be reluctant to share information for competitive or data privacy concerns.
3. Increasing richness of catalogue may raise concern with respect to patient privacy infringements and data security risks.
4. Technical difficulties with syncing and mitigation.
5. AAI aspects, logging of data access and content management
6. Ownership of the catalogue and responsibilities for technical and functional maintenance

Project team

WP3 consists of a core development team sponsored by BBMRI-NL. We consider the goals of BBMRI-NL and CTMM-TraIT in line with the Data4lifesciences goals and therefore list the complete pool of available people.

Core team:

- Salome Scholtens (UMCG, PI, *UMCG*)
- David van Enckevort (project lead, *BBMRI UMCG*)
- Marieke Bijlsma (data manager, *BBMRI UMCG*)
- Bart Charbon (software developer, *BBMRI UMCG*)
- Erik van Iperen (PhD student, *BBMRI*)
- Ricardo de Miranda Azevedo (Data manager, *BBMRI*)

In addition, the development is steered by a national catalogue working group. (Currently, a rate limiting factor is that there is no local funding for their activities). The national catalogue working group consists of the persons below. This group can be expanded during the project.

- Core team (see above)
- Jeneffer Lutomski (Radboud Biobank)
- Remco den Ouden (Radboud Biobank)
- Erik Flikkenschild (PSI)
- Aad van der Lugt (Erasmus MC)
- Peter Riegman (Erasmus MC)
- Martine Roos (UMCU)
- Fokke Terpstra (UMCU)
- Gerrit Meijer (NKI)
- Beatriz Carvalho (NKI)
- Jeroen Beliën (VUmc)
- Jan-Willem Boiten (Lygature)

Abbreviations

• BBMRI-ERIC	Biobanking and BioMolecular Research Infrastructure Europe
• BBMRI-NL	Biobanking and BioMolecular Research Infrastructure Netherlands
• BIOS	BBMRI-NL1.0 consortium to create a large-scale data infrastructure and to bring together BBMRI researchers focusing on integrative omics studies in Dutch Biobanks
• CTMM	The Center for Translational Molecular Medicine
• CTMM-TraIT	A Sustainable Infrastructure for Translational Biomedical Research
• FAIR	Findable, Accessible, Interoperable and re-useable see also http://datafairport.org/
• GoNL	Genome of the Netherlands
• PALGA	Pathologisch-Anatomisch Landelijk Geautomatiseerd Archief nationwide network and registry of histo- and cytopathology in the Netherlands
• PSI	Parelsnoer Institute
• UMC	University Medical Center