BBMRI.nl & Health-RI Expert Meeting:
How to make health(care) data available for research?

REPORT EXPERT MEETING #3

The third virtual BBMRI.nl and Health-RI expert meeting on practical approaches to make health(care) data available for research was held on the 26th of November 2020. This initiative is a collaborative effort of BBMRI.nl and Health-RI to implement a data-driven health research infrastructure for optimal access to knowledge, tools, facilities, health data and samples. This data-driven heath research infrastructure is fundamental in realising our ultimate goal of a learning healthcare system that enables sustainable and affordable personalized medicine and health. But how should this infrastructure be shaped? Which organisations are currently providing this service? And how do they provide access to their health(care) data?

The goal of this expert meeting series is to discuss the various practical approaches, and their strengths and weaknesses, to make different types of health(care) data available for scientific (re)use. During this meeting, Trynke de Jong presented the processes to find, access, request, share and link data within Lifelines. Pascal Suppers highlighted the experiences of DataHub. This meeting was attended by ~30 experts from various organisations and universities.

Introducing the new Lifelines data platform

The mission of Lifelines is to make data and samples from their longitudinal population cohort available worldwide for multidisciplinary research in the field of healthy ageing. Lifelines does not perform research but facilitates researchers in the field of healthy aging. Their cohort consists of ~167,000 persons from the three Northern provinces in the Netherlands: Drenthe, Groningen, and Friesland. This cohort is characterised by stable housing (low frequency of moving), ethnical uniformity with the vast majority being Caucasian, and a diverse socio-economic status. All participants signed an informed consent form, allowing Lifelines to give researchers access to the data and, in some cases, link the collected data to data deriving from other registries. The general assessment of the Lifelines cohort is divided in waves. Currently, Lifelines is in the third wave. Each wave includes a participant visit to perform measurements (e.g. anthropometry, anxiety depression, autofluorescence, and blood pressure) and a second visit to collect biological samples (fasting blood, DNA, 24-hour urine, faeces, etc.). In addition, questionnaires are being conducted during and between each assessment wave.

The data access procedure of Lifelines is depicted in figure 1. In order to receive access to the data, the scientific track record of the applicant as well as the quality of the submitted request are required to be positively reviewed. Upon approval, a data and material transfer agreement (DMTA) will be offered and a fee-for-service will be charged. Following, data will be released in a secure workspace or high-performance cluster (HPC). It is agreed upon to return sample results, derivates, and add-on data to Lifelines which can be made available to other researchers. Lifelines has an open protocol to allow for the collection of additional data or samples, at the initiative of researchers. These additional studies must be approved by

Lifelines and a medical ethical committee and may require additional informed consent. Approved additional studies vary strongly in scope and approach, from broadly distributed additional questionnaires (such as the repeated COVID-19 related questionnaires), to performing specialised measurements (imaging, eye exams, omics, pedometers, diaries) in a select target population within the cohort. Data resulting from these additional studies are integrated with the existing database. Furthermore, data linkage is being performed with numerous organisations, including statistics Netherlands (CBS), medical registries such as NIVEL, PALGA, IKNL, and environmental data. Linked datasets are typically created on demand per project and are not integrated in the database.

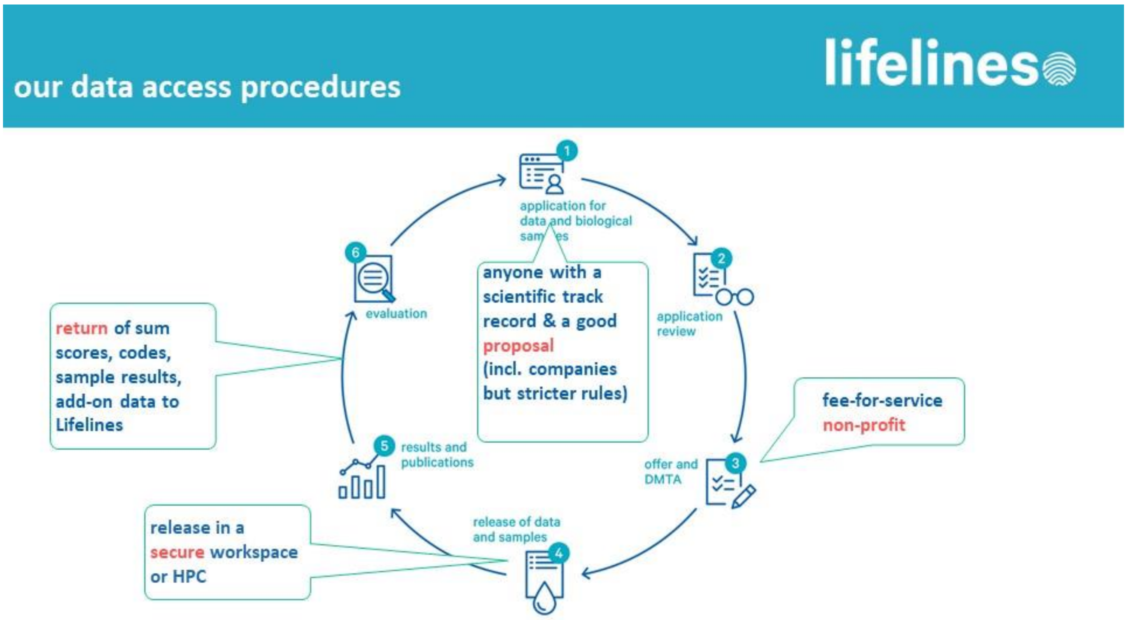
our data access procedures
lifelines
1
application for data and biological samples
anyone with a scientific track record & a good proposal (incl. companies but stricter rules)
2
application review
fee-for-service non-profit
3
offer and DMTA
4
release of data and samples
release in a secure workspace or HPC
5
results and publications
6
evaluation
return of sum scores, codes, sample results, add-on data to Lifelines

Figure 1. Data access procedure Lifelines

Lifelines used to store and release data in tables organised by measurement type, biosample type, or questionnaire theme. This approach resulted in challenges for the researchers, including 1) prior knowledge of details and exact contents of the dataset, 2) the precise protocol under which data is collected, 3) the story behind empty cells in the table, and 4) distinguishing non-responders from non-invited participants. Similarly, releasing data as data tables goes against the principle of data minimalization, and puts pressure on data managers to choose the right moment and structure for a data release. In order to overcome these challenges, Lifelines collaborated with Trivento to increase the resolution of the database by fragmenting the tables into individual data points enriched with basic metadata: "who" (e.g. participant ID and characteristics), "what" (e.g. variable information) and "when" (date and methodologic information). Data is then aggregated in 3 general who-what-when data tables and two cross tables, i.e. a full list of participants per variant (who x when) and a full list of variables per variant (when x what).

The www-metadata is fed into the updated catalogue (developed by Molgenis), enabling researchers to select specific age/sex/thematic subgroups and specific general/additional assessments. Data orders are submitted to the data platform, which automatically generates and organizes pseudonymized tables. The accessible dataset contains all basal metadata information. Additional aggregated metadata information is provided via the Lifelines Wiki to inform the researcher in more depth concerning the background and validity of used methods, specific protocols, additional studies and publications. If you have any questions or remarks, feel free to contact Trynke de Jong (t.r.de.jong@lifelines.nl).

DataHub, towards a FAIR future

DataHub is an institutional data broker that facilitates the findability and sharing of data. Based on the condition “structure early, share later” DataHub assists to develop an infrastructure that can be best described as a state-of-the-art GDPR compliant FAIR (Findable, Accessible, Interoperable, and Re-usable) data station. DataHub is made, managed and extended by the academic community, consisting of Maastricht University, the hospital Maastricht UMC+, and others such as Scannexus. Science is in a phase of rapid transition due to the rise of digital technology, open science and data protection regulations. DataHub offers support to individual research groups who want to stay up to date with the current developments in the field of digital technology. DataHub facilitates Open Science so that society as a whole can benefit from publicly funded research. In line with this, DataHub enables research data to go FAIR (Findable, Accessible, Interoperable, Re-usable): a bottom-up international approach for the practical implementation of the European Open Science Cloud (EOSC) as part of a global Internet of FAIR Data & Services. General Data Protection Regulation (GDPR) has triggered the developments of DataHub to support researchers to 1) specify how data should be used and protected, 2) to harmonize data protection laws across the EU Accountability, 3) establish a legal basis for data processing, 4) comply to privacy framework, 5) record all data processing activities, 6) establish data processing agreements, 7) appoint Data Protection Officers, and 8) comply to far-reaching data subject rights.

DataHub is an institutional data broker that aims to link community and technology to enable FAIR data. DataHub consist of a small multidisciplinary core team including research software engineers supplemented with embedded data stewards. This team possess the relevant domain knowledge: covering scientific knowledge from molecular to clinical research. DataHub facilitates data management of research conducted in the both hospitals and universities. DataHub provides use case-based development, meaning that the researchers are involved in the process of establishing the data management infrastructure in collaboration with the research software engineers. This is being achieved by an in-kind investment of the researcher on site with the DataHub data engineers alias research software engineers. The developed tools and services within DataHub are, where possible, open source to create a generic infrastructure facilitating extension and linking of the datasets.

Researchers are often required to write a Data Management Plan (DMP) when they apply for funding. DataHub offers DMP templates for researchers and a local DMP collaboration tool (based on DMP-online). DataHub supports researchers in the comprehension of policy and legislation by translating the GDPR into concrete instruction. This allows the researchers to be more in control on e.g. the risk for data leaks, data governance, compliance with (patient) privacy policies and de-identification guidelines.

Researchers can be granted easy access to a one stop shop Research Data Management. Data sets stored at the DataHub repository are FAIR by providing pseudonymization services via the Master Person Index (algorithm to connect human data). This is in collaboration with a trusted second party. The metadata in the DataHub catalogue is standardized using ontologies and a common data model, such as the OMOP common data model (OHDSI consortium). The data that is findable in the catalogue does not only include research data from Maastricht University and hospital Maastricht UMC+, but also other open data sources.

DataHub experiences challenges with regards to different types of data, varying in structure, personal, and meta data level. Applying structure to data is challenging but crucial for interoperability of data, using for example ontology SNOMED code. To achieve personal medicine, the FAIR principles need to be followed (see figure 2). Important while doing so, first start from enabling findability (F) of data to allow sound establishment of Accessibility (A), Interoperability (I), and Re-usability (R) at a later timepoint. Linking data catalogues would be a first step to achieve increased findability. If you have any questions or remarks, feel free to contact Pascal Suppers (p.suppers@maastrichtuniversity.nl).

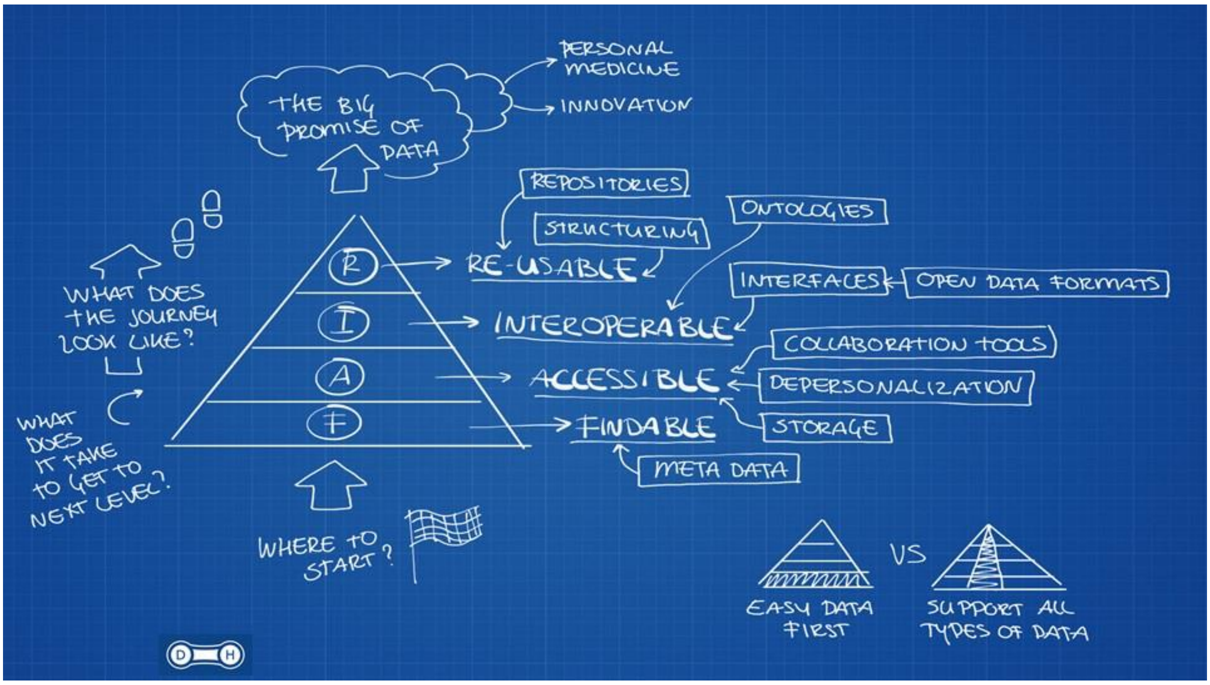

Figure 2. The big promise of data according DataHub

This series of expert meetings focused on providing examples of current data sharing procedures, available tools and challenges faced. Follow-up expert meetings will be organised during 2021 to discuss specific topics in more detail and to strengthen the data sharing community. Please contact Robin Verjans (robin.verjans@lygature.org) to inform us about the topics you prefer to discuss in-depth during future events or for any additional questions or remarks.