Working group mandate:

Data Lifecycle Management: Practices, Standards and Specifications for a Dataset Catalogue

Background

The implementation of a FAIR (Findability, Accessibility, Interoperability and Re-usability) data management strategy requires to manage the whole lifecycle of data in an integrated fashion, from the creation (generation or acquisition) up to the deletion (or long-term conservation) of the data. The data lifespan includes several processing steps, including data planning, data transfer, data transformation and management, data analysis and - optionally - data publication with or without access restriction depending on the nature of the data and the published study. Thus, aggregated statistical data may ultimately become fully public - as supplementary data in a published article - whereas participant-level data, collected in the context of personalized health research or a clinical trial, will have to be safely stored for at least 10 years. Whatever steps are punctuating the data lifecycle, the ability to carefully monitor those steps is paramount in FAIR.

Within this context, dataset catalogues, which keep track of the existence of datasets and under which conditions they can be accessed, are cornerstone of any data lifecycle management infrastructure. Such catalogues must provide effective instruments to search and browse available datasets. It must also provide a rich description of the data with a plurality of accurate and relevant attributes (e.g. contents of the data, size of the population), so that the researchers can initiate sufficiently informed dataset requests, thus avoiding wasting time with irrelevant dataset requests. In the context of secondary usage of datasets, these catalogues are needed to link back the enriched datasets to the original data.

Proposed mandate

The Working Group shall provide a survey of current practices in the field in Switzerland and identify relevant international examples of meta-data catalogues. It will explore what communities are active in the field and how these communities tackle the challenges associated with the design and development of such catalogues: meta-data description standards, best practices, references ontologies to describe these datasets, search instruments, etc. Further, the Working Group shall explore the different types of datasets likely to be described in such catalogues (research data, cohort data, public access data, restricted access data, registry data, data as stored in clinical data warehouses, etc.)

Furthermore, the Working Group shall explore the feasibility of designing a dataset catalogue to primarily find and access Swiss datasets. The specification of the catalogue should build upon existing guidelines, be compatible with the SPHN environment, principles, and efforts and integrated into the overall Swiss landscape. It will also pay attention to international initiatives (e.g. dbGap, EGA, GA4GH) in the domain.

Deliverables

– Mapping of current dataset catalogue landscape in Switzerland and outline of relevant international examples;

- Definition of roles and responsibilities of funders and regulators (e.g. SNSF, SwissMedic), data providers, PIs and academic institutions (data controllers), as well as infrastructure providers (e.g. DCC/BioMedIT network; data processors) over the full data lifecycle;
- A work plan (scope and focus, feasibility, timeline, work packages, resource requirements);

- Regarding **findability** of data:
  - Define a harmonized meta-data set for all relevant datasets (aligned with international initiatives) and data specific initiatives (in close collaboration with the cohort and registry data Working Group and other players);
  - Define a plan for federated or central meta-data search;
  - Define a plan to link enriched datasets using meta-data;
- Regarding **accessibility** of data:
  - Describe existing governance processes;
  - Develop a strategy for implementation of a central data request portal in accordance with the existing governance processes;
- Regarding **interoperability** of meta-data and related processes:
  - Identify processes, which should be harmonized to ensure interoperability of meta-data regarding both the biomedical contents of the datasets and the consent conditions;
  - Define a concept for national interoperability of those dataset catalogues (considering international standards, SPHN clinical and biological standards);
- Regarding **reuse** of data:
  - Provide a concept for a transparent management of re-usability restrictions (e.g. disease specific conditions, data storage duration, …)
  - Provide a concept for dataset versioning, as well as cross-references to external repositories (publications, other catalogues, …)
  - Provide a concept for a sustainable management (responsibilities, cost model, …) of the dataset catalogue according to different storage duration (active data up to publication, publication + 2-5 years, long term storage).

Proposed work plan:

A series of 4-6 meetings is arranged with the Working Group. The meeting schedule is according to the availability of the members. During the meetings, the draft strategy is discussed and further developed. The meeting documentation should be sent at least 1 week prior to the meeting.

Financial needs for the working group:

Reimbursement of travel expenses for Swiss participants	
Reimbursement of travel expenses for international experts	
Refreshments	
Total	**CHF 20’000.-**

Proposed composition of the working group

Members from infrastructure and data providers:

- Thierry Sengstag, sciCORE, University of Basel
- Heinz Stockinger, Core-IT, SIB
- Bernd Rinn, SIS, ETHZ
- Cornelia Kruschel, USZ

Member of a funding institution:

- SNSF (name to be determined)

Members from research:

- Andre Kahles (replacing Gunnar Rätsch), ETHZ
- Constantin Sluka (replacing Milica Marcovic), USB
- Rémy Bruggmann, University of Berne

Representative of swissuniversities / Swiss National Open Science Strategy:

- Name to be determined

Members from international institutions involved in the maintaining of existing catalogues, to be consulted as external advisors:

- Helen Parkinson or Thomas Keane, EGA, EBI-EMBL
- Niclas Jareborg, SciLifeLab, National Bioinformatics Infrastructure Sweden

Chairperson: Patrick Ruch, HES-SO, SIB
Coordination: PHI group