

## The data life cycle in 1+MG

v1.0 from 4 August 2022

### Definitions

The documents are built on the [1+MG Glossary](#). In particular, the following expressions are used:

- **Metadata:** a set of data that describes and gives information about other data. Metadata do not include any data that are processed to produce any results of the data use such as health or genetic data. Metadata can be personal data where they contain information on the subject level (e.g. consent decisions of an individual data subject)
- **Data collections:** Data collections are data that come from the same collection context from the same controller(s) and can be characterised with non-personal metadata. A data collection can e.g. refer to data of a cohort or a public registry.
- **Records:** Data related to individual data subjects
- **Data set:** Data that are grouped in a certain context, e.g. for a user's access.
- **Enriched data:** data where additional information is added, e.g. annotations
- **Derived data:** data that have been created by alteration of the original data (e.g. through data curation)

### The data journey in 1+MG

The following stages are considered for the data journey.

1. Data preparation: Pre-processing to agreed standards, annotation with metadata etc
2. Data inclusion: Physical transfer of data incl., legal transfer of data to 1+MG to enable visibility in data catalogue
3. Data storage and management: Including GDPR Compliant processing environment, data versioning etc
4. Data discovery: Discovery of data using GDPR compliant APIs e.g., Beacon
5. Data access: A mechanism(s) by which the data controller can authorise access to select dataset(s)
6. Data use: Data processing for the approved purposes to obtain a result
7. Data archiving for the approved purposes (where necessary for a respective purpose)

*Please note that the current data governance for research use on which the DPbDD recommendations are based do not yet cover access by users outside the European Economic Area (EEA). In addition, the data governance for access in the context of rare diseases, which falls between research and healthcare, is also not yet decided.*

### Data's life in 1+MG

Data included in 1+MG have already been generated in a different context (primary use) and are subsequently brought into 1+MG by the data holder. It is the responsibility of the data holder that the data are characterised with relevant metadata and curated into a data model accepted by 1+MG. 1+MG and potentially national / local structures will support data holders. Data are uploaded by a data holder, jointly with metadata. The data are ingested into the system and relevant metadata made available in the data catalogue. A user may choose from the data catalogue relevant data collections or, after registration for research use, even individual records across different collections and countries. Based on the

selection, the user launches an access request. The signing official of the user's institution must confirm this request unless the user is in a list of mandated users in the system, which flags that the researcher has a permanently confirmed right in the home institutions to apply for access. In the case of healthcare professionals, proof of their profession may also be sufficient. The information on the request with all relevant information is sent to the 1+MG central office of the central 1+MG Data Access Committee (central DAC) and the relevant National Coordination Points. The central office communicates to the relevant National Coordination Points the justified access decision of the central DAC. In case of research access, the National Coordination Point (after consultation of relevant stakeholders, where applicable) may or may not communicate a veto in a certain time frame. Positive feedback can be sent as well to shorten the response time for the access request. In case of approval and a signed contract, the user receives access to the data set that was applied for. This may include subsets of records according to data types that can be chosen and therefore a definition of a "minimum data unit" that can be managed. The user may need an easy to use, intuitive interface for querying data and tracking their applications, data that they have been granted access and expiration date, how to extend them etc.. Other users need the possibility to run more complex analyses on the data; some may even bring in their own data and /or own algorithm. This may require a more complex compute environment. Data may change over time because of new data points added (additional data types, additional time points) or data subjects' requests (deletion, rectification). In addition, data use may lead to enriched or derived data also available through 1+MG. The users must be able to update datasets but also to request a stable dataset. For some use, archiving of the analysis and the possibility to re-run the analysis again up to 10 years later is required. A data versioning (release) mechanism and tools are therefore needed. This should also include the possibility to identify when dataset versions can be deleted because there is no longer a need to retain them.