

README: Version 1 Release of Minimal Data Sets for the 1+MG/GDI Project







Overview

This document serves as the README for the version 1 release of minimal data sets for the 1+MG/GDI project, dated February 2025. This release includes a set of spreadsheets that contain items as decided by the 1+MG WG3 & GDI Task 8.2 group.

Contents

This release consists of 5 spreadsheets, each containing structured data entries based on ontological terms. The primary goal of this release is to ensure that values are properly referenced using established ontologies and that missing terms are handled appropriately.

The spreadsheets are:

-  Sample Class Version 1
-  Subject and Diagnosis Class Version 1
-  Treatment Class Version 1
-  Environmental Exposure and Biomarker Class Version 1
-  Sequence Class Version 1
-  Metadata Submission Version 1

Next all these structured data entries, in the form of classes have a certain relation to each-other (called cardinalities). These relations are described/pictured via a UML-diagram.

The UML-diagram can be found here:

- [GDI_diagram.drawio](#)
-  GDI_diagram.pdf

Classes, Properties and Values

- Ultimately the goal is to arrive at a full knowledge graph. For now most are composed of ontological terms, which are provided as links within the spreadsheets.
 - Concepts are defined in column
 - description/label: B (Suggested Union Item)
 - definition : C (Suggested_Union_Proposed_Definition)

- codesystem/code: I (Link to terminology/ontology that defines item)
- Values are defined in column (in the end Column D and J if value lists will be the same)
 - description/label: D (Suggested_Union_Proposed_Values)
 - codesystem/code: J (Link to terminology/ontology that defines value(s)/valuelist(s))
- If values are missing proper ontological terms, they will be temporarily moved to the [FAIR Genomes semantic model](#) to create appropriate namespaces.
- If a suitable ontological term is identified in the future, the FAIR Genomes namespace will be deprecated and will point to the newly established namespace.

Implementation of Lists

- Lists should ideally consist of **ontological terms**.
- Some lists contain terms for which no appropriate ontology terms were found. In these cases, the lists were moved to FAIR Genomes, where fitting namespaces were created.
- Lists that contain a mix of known ontological terms and missing terms will be entirely moved to FAIR Genomes. However, terms that already have established ontological references remain linked to them.
- Values in a list should always be implemented using the **original IRI** whenever possible.
- If necessary for implementation convenience, all lists could be moved to FAIR Genomes.
- For mandatory items in all classes with a list, also the HL7 NULL flavors will be valid.

Key Guidelines for Implementation:

1. Even if a list is entirely present in FAIR Genomes, do not merely reference the list in the FAIR Genomes repository. Instead, list individual terms in the original ontology.
2. The **FAIR Genomes namespace is temporary** – if a better ontological term is identified in the future, it will replace the FAIR Genomes namespace.
3. The FAIR Genomes namespace should **not be removed** but should instead point to the new namespace.