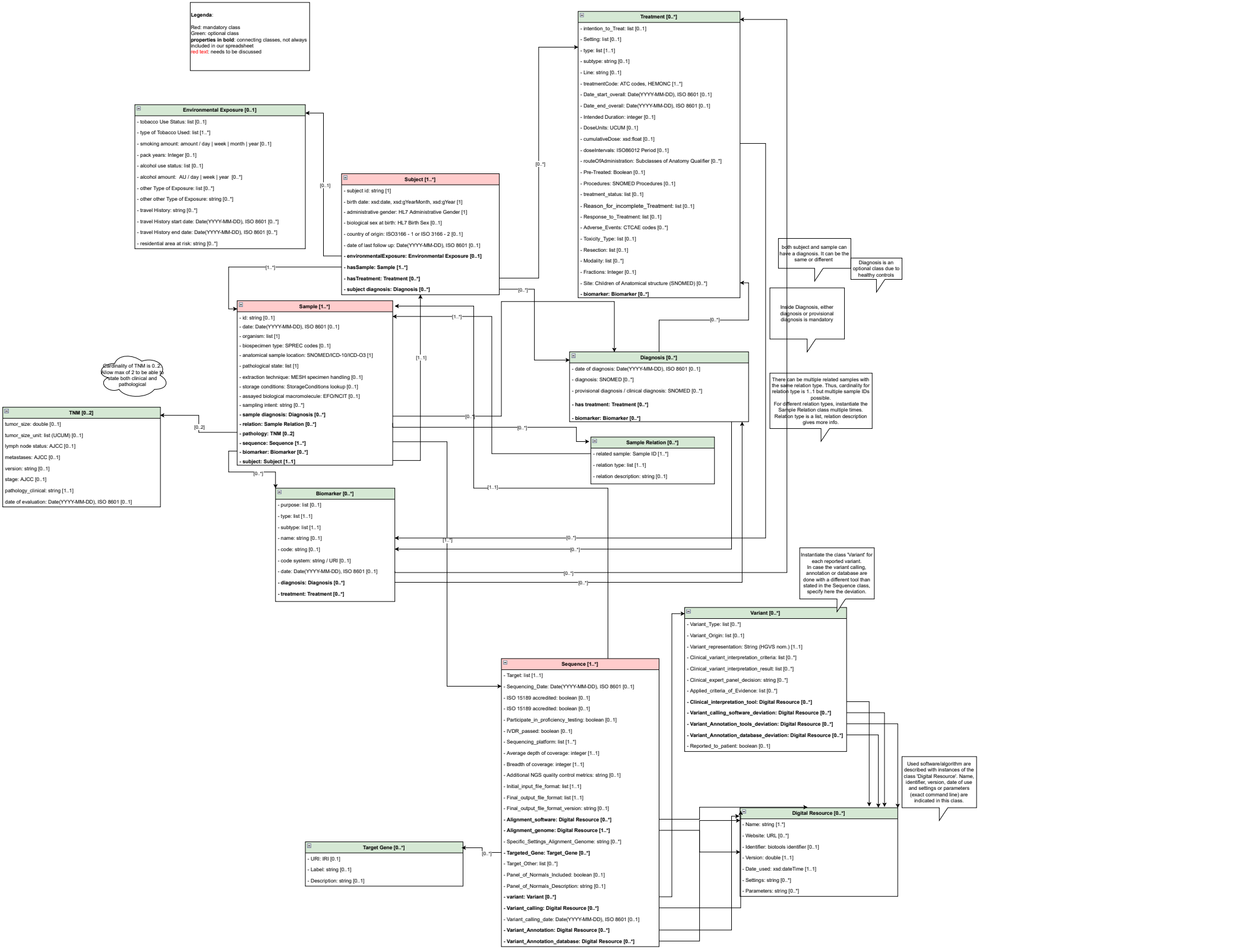
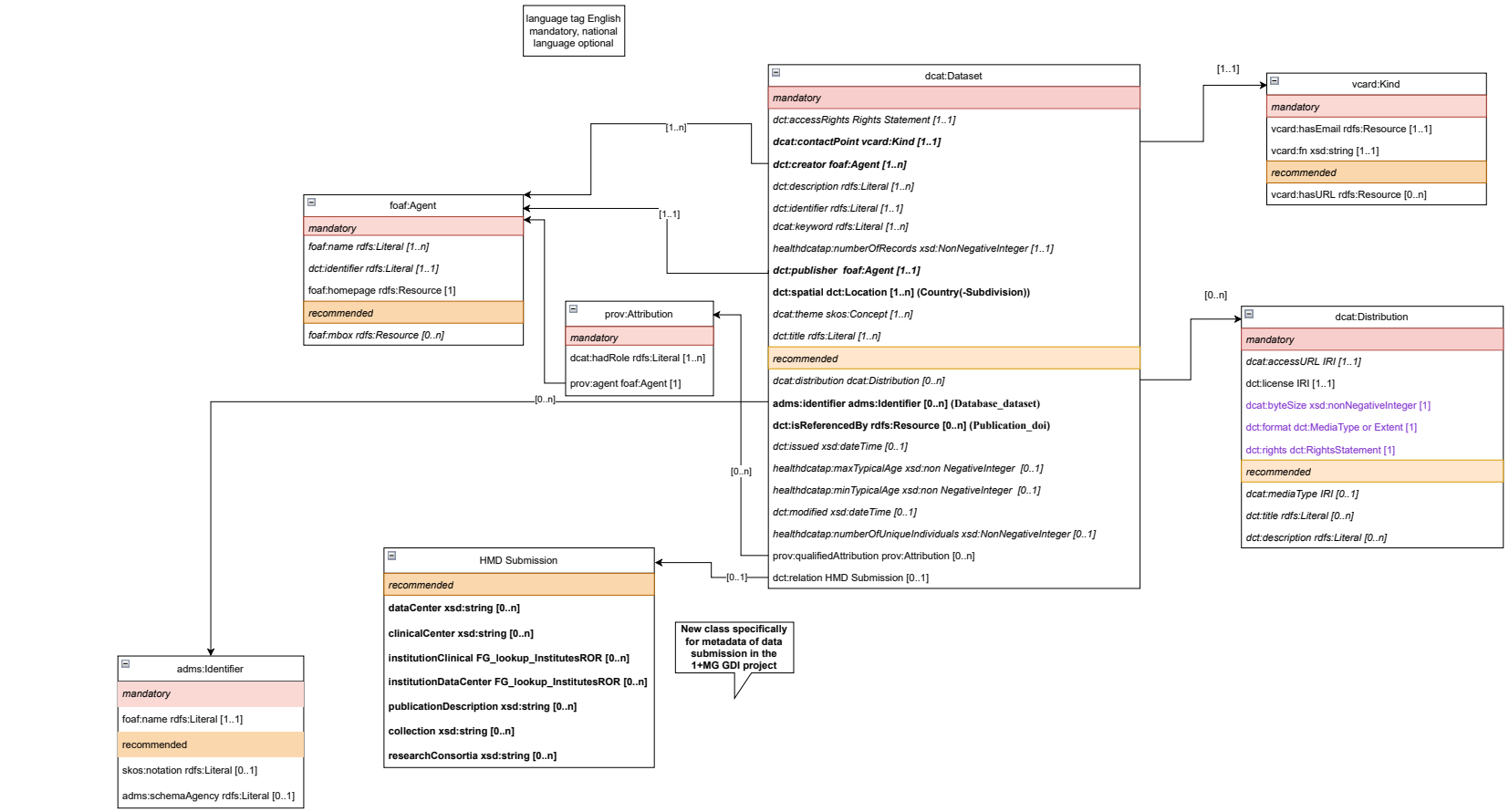
Tab 2: UML Diagram 1+MG GDI

Proposed diagram for the 1+MG GDI T8.2 minimal dataset Classes (incl. their respective properties) and their relations are depicted.
Modelling based on combining efforts/models from the different use cases (Cancer, Rare Diseases, Infectious Diseases).

Tab 3: Metadata

Proposed diagram for metadata of data submission. This work is based on the metadata submission class in the data model, GDI MS8 metadata items, and the core metadata schema of Health-RI (NL), which currently includes the (draft) of HealthDCAT-AP (developed in TEHDAS2 program; version of 16-12-2025 was used), see legend.


Legenda:
Red: mandatory class
Green: optional class
properties in bold: connecting classes, not always included in our spreadsheet
red text: needs to be discussed
Environmental Exposure [0..1]
- tobacco Use Status: list [0..1]
- type of Tobacco Used: list [1..*]
- smoking amount: amount / day | week | month | year [0..1]
- pack years: Integer [0..1]
- alcohol use status: list [0..1]
- alcohol amount: AU / day | week | year [0..*]
- other Type of Exposure: list [0..*]
- other other Type of Exposure: string [0..*]
- travel History: string [0..*]
- travel History start date: Date(YYYY-MM-DD), ISO 8601 [0..*]
- travel History end date: Date(YYYY-MM-DD), ISO 8601 [0..*]
- residential area at risk: string [0..*]
Subject [1..*]
- subject id: string [1]
- birth date: xsd:date, xsd:gYearMonth, xsd:gYear [1]
- administrative gender: HL7 Administrative Gender [1]
- biological sex at birth: HL7 Birth Sex [0..1]
- country of origin: ISO3166 - 1 or ISO 3166 - 2 [0..1]
- date of last follow up: Date(YYYY-MM-DD), ISO 8601 [0..1]
- environmentalExposure: Environmental Exposure [0..1]
- hasSample: Sample [1..*]
- hasTreatment: Treatment [0..*]
- subject diagnosis: Diagnosis [0..*]
Treatment [0..*]
- intention_to_Treat: list [0..1]
- Setting: list [0..1]
- type: list [1..1]
- subtype: string [0..1]
- Line: string [0..1]
- treatmentCode: ATC codes, HEMONC [1..*]
- Date_start_overall: Date(YYYY-MM-DD), ISO 8601 [0..1]
- Date_end_overall: Date(YYYY-MM-DD), ISO 8601 [0..1]
- Intended Duration: integer [0..1]
- DoseUnits: UCUM [0..1]
- cumulativeDose: xsd:float [0..1]
- doseIntervals: ISO86012 Period [0..1]
- routeOfAdministration: Subclasses of Anatomy Qualifier [0..*]
- Pre-Treated: Boolean [0..1]
- Procedures: SNOMED Procedures [0..1]
- treatment_status: list [0..1]
- Reason_for_incomplete_Treatment: list [0..1]
- Response_to_Treatment: list [0..1]
- Adverse_Events: CTCAE codes [0..*]
- Toxicity_Type: list [0..1]
- Resection: list [0..1]
- Modality: list [0..*]
- Fractions: Integer [0..1]
- Site: Children of Anatomical structure (SNOMED) [0..*]
- biomarker: Biomarker [0..*]
both subject and sample can have a diagnosis. It can be the same or different
Diagnosis is an optional class due to healthy controls
Inside Diagnosis, either diagnosis or provisional diagnosis is mandatory
Sample [1..*]
- id: string [0..1]
- date: Date(YYYY-MM-DD), ISO 8601 [0..1]
- organism: list [1]
- biospecimen type: SPREC codes [0..1]
- anatomical sample location: SNOMED/ICD-10/ICD-O3 [1]
- pathological state: list [1]
- extraction technique: MESH specimen handling [0..1]
- storage conditions: StorageConditions lookup [0..1]
- assayed biological macromolecule: EFO/NCIT [0..1]
- sampling intent: string [0..*]
- sample diagnosis: Diagnosis [0..*]
- relation: Sample Relation [0..*]
- pathology: TNM [0..2]
- sequence: Sequence [1..*]
- biomarker: Biomarker [0..*]
- subject: Subject [1..1]
Cardinality of TNM is 0..2, allow max of 2 to be able to state both clinical and pathological
Diagnosis [0..*]
- date of diagnosis: Date(YYYY-MM-DD), ISO 8601 [0..1]
- diagnosis: SNOMED [0..*]
- provisional diagnosis / clinical diagnosis: SNOMED [0..*]
- has treatment: Treatment [0..*]
- biomarker: Biomarker [0..*]
There can be multiple related samples with the same relation type. Thus, cardinality for relation type is 1..1 but multiple sample IDs possible.
For different relation types, instantiate the Sample Relation class multiple times.
Relation type is a list, relation description gives more info.
TNM [0..2]
tumor_size: double [0..1]
tumor_size_unit: list (UCUM) [0..1]
lymph node status: AJCC [0..1]
metastases: AJCC [0..1]
version: string [0..1]
stage: AJCC [0..1]
pathology_clinical: string [1..1]
date of evaluation: Date(YYYY-MM-DD), ISO 8601 [0..1]
Sample Relation [0..*]
- related sample: Sample ID [1..*]
- relation type: list [1..1]
- relation description: string [0..1]
Biomarker [0..*]
- purpose: list [0..1]
- type: list [1..1]
- subtype: list [1..1]
- name: string [0..1]
- code: string [0..1]
- code system: string / URI [0..1]
- date: Date(YYYY-MM-DD), ISO 8601 [0..1]
- diagnosis: Diagnosis [0..*]
- treatment: Treatment [0..*]
Instantiate the class 'Variant' for each reported variant.
In case the variant calling, annotation or database are done with a different tool than stated in the Sequence class, specify here the deviation.
Variant [0..*]
- Variant_Type: list [0..*]
- Variant_Origin: list [0..1]
- Variant_representation: String (HGVS nom.) [1..1]
- Clinical_variant_interpretation_criteria: list [0..*]
- Clinical_variant_interpretation_result: list [0..*]
- Clinical_expert_panel_decision: string [0..*]
- Applied_criteria_of_Evidence: list [0..*]
- Clinical_interpretation_tool: Digital Resource [0..*]
- Variant_calling_software_deviation: Digital Resource [0..*]
- Variant_Annotation_tools_deviation: Digital Resource [0..*]
- Variant_Annotation_database_deviation: Digital Resource [0..*]
- Reported_to_patient: boolean [0..1]
Sequence [1..*]
- Target: list [1..1]
- Sequencing_Date: Date(YYYY-MM-DD), ISO 8601 [0..1]
- ISO 15189 accredited: boolean [0..1]
- ISO 15189 accredited: boolean [0..1]
- Participate_in_proficiency_testing: boolean [0..1]
- IVDR_passed: boolean [0..1]
- Sequencing_platform: list [1..*]
- Average depth of coverage: integer [1..1]
- Breadth of coverage: integer [1..1]
- Additional NGS quality control metrics: string [0..1]
- Initial_input_file_format: list [1..1]
- Final_output_file_format: list [1..1]
- Final_output_file_format_version: string [0..1]
- Alignment_software: Digital Resource [0..*]
- Alignment_genome: Digital Resource [1..*]
- Specific_Settings_Alignment_Genome: string [0..*]
- Targeted_Gene: Target_Gene [0..*]
- Target_Other: list [0..*]
- Panel_of_Normals_Included: boolean [0..1]
- Panel_of_Normals_Description: string [0..1]
- variant: Variant [0..*]
- Variant_calling: Digital Resource [0..*]
- Variant_calling_date: Date(YYYY-MM-DD), ISO 8601 [0..1]
- Variant_Annotation: Digital Resource [0..*]
- Variant_Annotation_database: Digital Resource [0..*]
Used software/algorithm are described with instances of the class 'Digital Resource'. Name, identifier, version, date of use and settings or parameters (exact command line) are indicated in this class.
Digital Resource [0..*]
- Name: string [1..*]
- Website: URL [0..*]
- Identifier: biotools identifier [0..1]
- Version: double [1..1]
- Date_used: xsd:dateTime [1..1]
- Settings: string [0..*]
- Parameters: string [0..*]
Target Gene [0..*]
- URI: IRI [0..1]
- Label: string [0..1]
- Description: string [0..1]



Legenda

Italics - Included in MS8 deliverable
Purple text - Mandatory field in HRI version 2, but not in GDI MS8 deliverable
Bold - Part of Metadata Submission class

Based on the HealthDCAT-AP version from date 16-12-2024