BS ISO 4454:2022

BSI Standards Publication

Genomics informatics — Phenopackets: A format for phenotypic data exchange

Documents
Document Sharing Hub

Full document details: BS ISO 4454:2022

National foreword

This British Standard is the UK implementation of ISO 4454:2022.

The UK participation in its preparation was entrusted to Technical Committee IST/35, Health informatics.

A list of organizations represented on this committee can be obtained on request to its committee manager.

Contractual and legal considerations

This publication has been prepared in good faith, however no representation, warranty, assurance or undertaking (express or implied) is or will be made, and no responsibility or liability is or will be accepted by BSI in relation to the adequacy, accuracy, completeness or reasonableness of this publication. All and any such responsibility and liability is expressly disclaimed to the full extent permitted by the law.

This publication is provided as is, and is to be used at the recipient's own risk.

The recipient is advised to consider seeking professional guidance with respect to its use of this publication.

This publication is not intended to constitute a contract. Users are responsible for its correct application.

© The British Standards Institution 2022
Published by BSI Standards Limited 2022

ISBN 978 0 539 14689 9

ICS 35.240.80

Compliance with a British Standard cannot confer immunity from legal obligations.

This British Standard was published under the authority of the Standards Policy and Strategy Committee on 31 July 2022.

Amendments/corrigenda issued since publication

Date	Text affected

Documents
Document Sharing Hub

Full document details: BS ISO 4454:2022

INTERNATIONAL STANDARD

ISO 4454

First edition
2022-07

Genomics informatics — Phenopackets: A format for phenotypic data exchange

Reference number
ISO 4454:2022(E)

Documents
Document Sharing Hub

© ISO 2022

Full document details: BS ISO 4454:2022

COPYRIGHT PROTECTED DOCUMENT

© ISO 2022

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub

Full document details: BS ISO 4454:2022

Contents

Page

Foreword ... **viii**
Introduction ... **ix**
1 Scope ... **1**
2 Normative references ... **1**
3 Terms and definitions ... **1**
4 Abbreviated terms ... **4**
5 Phenopackets Schema and Requirements ... **5**
5.1 Phenopacket Schema ... 5
5.2 Requirement Levels ... 11
5.2.1 General ... 11
5.2.2 Multiplicity ... 11
5.3 Ontology Use ... 12
6 Phenopacket Schema Top-Level Elements ... **12**
6.1 Phenopacket ... 12
6.1.1 General ... 12
6.1.2 id ... 12
6.1.3 subject ... 12
6.1.4 phenotypic_features ... 13
6.1.5 measurements ... 13
6.1.6 biosamples ... 13
6.1.7 interpretation ... 13
6.1.8 diseases ... 13
6.1.9 medical_actions ... 13
6.1.10 files ... 13
6.1.11 meta_data ... 13
6.2 Family ... 13
6.2.1 General ... 13
6.2.2 id ... 14
6.2.3 proband ... 14
6.2.4 relatives ... 14
6.2.5 pedigree ... 14
6.2.6 files ... 14
6.2.7 meta_data ... 14
6.3 Cohort ... 15
6.3.1 General ... 15
6.3.2 id ... 15
6.3.3 description ... 15
6.3.4 members ... 15
6.3.5 files ... 15
6.3.6 meta_data ... 15
7 Phenopackets Building Blocks ... **15**
7.1 General ... 15
7.2 Age ... 16
7.3 AgeRange ... 16
7.4 Biosample ... 16
7.4.1 General ... 16
7.4.2 id ... 18
7.4.3 individual_id ... 18
7.4.4 derived_from_id ... 18
7.4.5 description ... 18
7.4.6 sampled_tissue ... 18
7.4.7 sample_type ... 18

© ISO 2022 – All rights reserved

AVA Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

7.4.8 phenotypic_features 18
7.4.9 measurements 18
7.4.10 taxonomy 19
7.4.11 time_at_collection 19
7.4.12 histological_diagnosis 19
7.4.13 tumor_progression 19
7.4.14 tumor_grade 19
7.4.15 pathological_stage 19
7.4.16 pathological_tnm_finding 19
7.4.17 diagnostic_markers 19
7.4.18 procedure 19
7.4.19 files 19
7.4.20 material_sample 20
7.4.21 sample_processing 20
7.4.22 sample_storage 20
7.5 ComplexValue 20
7.5.1 General 20
7.5.2 typed_quantities 20
7.6 Disease 21
7.6.1 General 21
7.6.2 term 21
7.6.3 excluded 21
7.6.4 onset 22
7.6.5 resolution 22
7.6.6 disease_stage 22
7.6.7 clinical_tnm_finding 22
7.6.8 primary_site 22
7.6.9 laterality 22
7.7 DoseInterval 22
7.7.1 General 22
7.7.2 quantity 23
7.7.3 schedule_frequency 23
7.7.4 interval 23
7.8 DrugType 23
7.8.1 General 23
7.9 Evidence 24
7.9.1 General 24
7.9.2 evidence_code 24
7.9.3 Reference 24
7.10 ExternalReference 24
7.10.1 General 24
7.10.2 id 25
7.10.3 reference 25
7.10.4 description 25
7.11 File 25
7.11.1 General 25
7.11.2 uri 26
7.11.3 individual_to_file_identifiers 26
7.11.4 file_attributes 26
7.12 GeneDescriptor 26
7.12.1 General 26
7.12.2 value_id 27
7.12.3 symbol 28
7.12.4 description 28
7.12.5 alternate_ids 28
7.12.6 alternate_symbols 28
7.12.7 xrefs 28
7.13 GenomicInterpretation 28

© ISO 2022 – All rights reserved
AVA Documents Document Sharing Hub
Full document details: BS ISO 4454:2022

7.13.1 General 28
7.13.2 subject_or_biosample_id 29
7.13.3 interpretation_status 29
7.13.4 call 29
7.14 GestationalAge 29
7.14.1 General 29
7.15 Individual 30
7.15.1 General 30
7.15.2 id 30
7.15.3 alternate_ids 31
7.15.4 date_of_birth 31
7.15.5 time_at_last_encounter 31
7.15.6 vital_status 32
7.15.7 sex 32
7.15.8 karyotypic_sex 32
7.15.9 gender 32
7.15.10 taxonomy 32
7.16 Interpretation 32
7.16.1 General 32
7.16.2 id 33
7.16.3 progress_status 33
7.16.4 diagnosis 34
7.16.5 summary 34
7.17 KaryotypicSex 34
7.18 Measurement 35
7.18.1 General 35
7.18.2 description 36
7.18.3 assay 36
7.18.4 value 36
7.18.5 time_observed 36
7.18.6 procedure 36
7.19 MedicalAction 36
7.19.1 General 36
7.19.2 action 37
7.19.3 treatment_target 37
7.19.4 treatment_intent 37
7.19.5 response_to_treatment 37
7.19.6 adverse_events 37
7.19.7 treatment_termination_reason 37
7.20 MetaData 37
7.20.1 General 37
7.20.2 created 38
7.20.3 created_by 38
7.20.4 submitted_by 38
7.20.5 resources 38
7.20.6 updates 39
7.20.7 phenopacket_schema_version 39
7.20.8 external_references 39
7.21 OntologyClass 39
7.21.1 General 39
7.21.2 id 39
7.21.3 label 39
7.22 Pedigree 40
7.22.1 General 40
7.22.2 persons 40
7.23 PhenotypicFeature 42
7.23.1 General 42
7.23.2 description 42

© ISO 2022 – All rights reserved

AVA Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

7.23.3 type 42
7.23.4 excluded 43
7.23.5 severity 43
7.23.6 modifiers 43
7.23.7 onset 43
7.23.8 resolution 43
7.23.9 evidence 43
7.24 Procedure 43
7.24.1 General 43
7.24.2 code 44
7.24.3 body site 44
7.24.4 performed 44
7.25 Quantity 44
7.25.1 General 44
7.25.2 unit 45
7.25.3 value 45
7.25.4 reference_range 45
7.26 RadiationTherapy 45
7.26.1 General 45
7.26.2 modality 45
7.26.3 body_site 45
7.26.4 dosage 46
7.26.5 fractions 46
7.27 ReferenceRange 46
7.27.1 General 46
7.27.2 unit 46
7.27.3 low 46
7.27.4 high 46
7.28 Resource 46
7.28.1 General 46
7.28.2 id 47
7.28.3 name 48
7.28.4 url 48
7.28.5 version 48
7.28.6 namespace_prefix 48
7.28.7 iriPrefix 48
7.28.8 CURIE 48
7.28.9 Identifier resolution 49
7.29 Sex 49
7.30 TherapeuticRegimen 49
7.30.1 General 49
7.30.2 identifier 50
7.30.3 start_time 50
7.30.4 end_time 50
7.30.5 regimen_status 50
7.31 TimeElement 50
7.31.1 General 50
7.31.2 gestational_age 51
7.31.3 age 51
7.31.4 age_range 51
7.31.5 ontology_class 51
7.31.6 timestamp 51
7.31.7 interval 51
7.32 TimeInterval 51
7.32.1 General 51
7.32.2 start 52
7.32.3 end 52
7.33 Timestamp 52

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

7.34 Treatment 52
7.34.1 General 52
7.34.2 agent 54
7.34.3 route_of_administration 54
7.34.4 dose_intervals 54
7.34.5 drug_type 54
7.34.6 cumulative_dose 54
7.35 Update 54
7.35.1 General 54
7.35.2 timestamp 55
7.35.3 updated_by 55
7.35.4 comment 55
7.36 Value 55
7.36.1 General 55
7.36.2 value 55
7.37 VariationDescriptor 55
7.37.1 General 55
7.37.2 id 58
7.37.3 variation 58
7.37.4 label 58
7.37.5 description 58
7.37.6 gene_context 58
7.37.7 expression 58
7.37.8 vcf_record 58
7.37.9 xrefs 59
7.37.10 alternate_labels 59
7.37.11 extensions 59
7.37.12 molecule_context 59
7.37.13 structural_type 59
7.37.14 vrs_ref_allele_seq 60
7.37.15 allelic_state 60
7.38 VariantInterpretation 60
7.38.1 General 60
7.38.2 acmg_pathogenicity_classification 60
7.38.3 therapeutic_actionability 61
7.38.4 variant 61
7.39 VitalStatus 61
7.39.1 General 61
7.39.2 status 61
7.39.3 time_of_death 62
7.39.4 cause_of_death 62
7.39.5 survival_time_in_days 62

Annex A (informative) **Working with Phenopackets** **63**

Annex B (informative) **Ontologies** **71**

Annex C (informative) **Rare Disease Example** **74**

Annex D (informative) **Cancer Example** **77**

Annex E (informative) **COVID-19 Example** **81**

Bibliography **86**

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 215, *Health informatics*, Subcommittee SC 1, *Genomics informatics*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

Introduction

While great strides have been made in exchange formats for sequence and variation data (e.g. Variant Call Format), the majority of genotype formats do not include a means to share corresponding phenotypic (e.g. observable characteristics, signs/symptoms of disease) information. While some genomic databases have defined their own formats for representing phenotypic information, the lack of uniformity amongst these organizations hinders communication and limits the ability to perform analysis across organizations. For individuals with rare and undiagnosed disease, broad adoption and utilization of uniform, machine-readable, phenotypic descriptions could improve the speed and accuracy of diagnosis by promoting quicker, more comprehensive and cost-effective information acquisition and exchange relevant for research and medical care.

Phenotypic abnormalities of individuals are currently described in diverse places in diverse formats, such as journal/publications databases, laboratory systems, patient registries, health records, and even in social media. The structure of the data in the phenopackets exchange standard will be optimized for integration and efficient data flow across these distributed contexts. Increasing the volume of computable data across a diversity of systems will support large-scale computational disease analysis of combined genotype and phenotype data. Studies of well over 100 000 patients are thought to be required to effectively assess the role of rare variation in common disease or to discover the genomic basis for a substantial portion of diseases. Phenopackets can help integrate geographically distributed cases to build such virtual cohorts and remove the time burden on resources that need to integrate information manually.

Medical coding systems and clinical exchange standards have not to date included rich phenotypic descriptions, as they are largely focused on supporting billing and clinical encounter documentation, rather than the documenting and sharing of the biologically relevant phenotypic information needed for computational use, mechanism discovery, and precision classification. From a clinical perspective, the integration of a standard for phenotypic description and exchange into and out of EHRs would improve disease diagnosis and management, especially for genomic health and precision medicine applications.

Phenopackets enable clinicians, biologists, and disease and drug researchers to build more complete models of disease. It is designed to encourage wide adoption and synergy between the people, organizations and systems that comprise the joint effort to address human disease and biological understanding. The phenopacket proposed in this document is designed to support deep phenotyping, a process wherein individual components of each phenotype are observed and documented. The phenotypes can be constitutional or those related to a sample (such as from a biopsy).

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

Documents
Document Sharing Hub

Genomics informatics — Phenopackets: A format for phenotypic data exchange

1 Scope

This document specifies a uniform, machine-readable, phenotypic description of an individual, patient or sample in the context of rare disease, common/complex disease or cancer.

It is applicable to academic, clinical and commercial research, as well as clinical diagnostics. While intended for human data collection, it can be used in other areas (e.g. mouse research). It does not define the phenotypic information that needs to be collected for a particular use but represents that information in an appropriately descriptive manner that allows it to be computationally exchanged between systems.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 8601 (all parts), *Date and time — Representations for information interchange*

3 Terms and definitions

For the purposes of this document, the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
biosample
unit of biological material from which the substrate for analysis is extracted to support the assessment, diagnosis, treatment, mitigation or prevention of a disease, disorder, abnormal physical state or its symptoms

3.2
boolean
data type having two values: one and zero (which are equivalent to true and false)

[SOURCE: ISO 2146:2010, 4.6.1]

3.3
CURIE
compact URI
generic, abbreviated syntax for expressing *uniform resource identifiers* (3.22)

3.4
deletion
variation in which a part of a chromosome or sequence of DNA is lost relative to a *reference sequence* (3.17)

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

3.5
DNA sequence
order of *nucleotide bases* (3.10) (adenine, guanine, cytosine and thymine) in a molecule of DNA

3.6
exome sequencing
technique for sequencing the protein-coding genes in a genome

[SOURCE: ISO/TS 20428:2017, 3.38, modified — "whole" removed from preferred term.]

3.7
gene
basic unit of hereditary information composed of chains of nucleotides in specific sequences that encodes a protein or protein subunit

[SOURCE: ISO 11238:2018, 3.29]

3.8
gestational age
menstrual age
time elapsed between the first day of the last normal menstrual period and the day of delivery

Note 1 to entry: The first day of the last menstrual period occurs approximately 2 weeks before ovulation and approximately 3 weeks before implantation of the blastocyst. Because most women know when their last period began but not when ovulation occurred, this definition traditionally has been used when estimating the expected date of delivery. In contrast, chronological age (or postnatal age) is the time elapsed after birth.

3.9
insertion
addition of one or more *nucleotide base pairs* (3.10) into a *DNA sequence* (3.5)

[SOURCE: ISO/TS 20428:2017, 3.19]

3.10
nucleotide base
nucleotide base pair
monomer of a nucleic acid polymer such as DNA or RNA

Note 1 to entry: Nucleotides are denoted as letters ('A' for adenine; 'C' for cytosine; 'G' for guanine; 'T' for thymine that only occurs in DNA; and 'U' for uracil that only occurs in RNA). The chemical formula for a specific DNA or RNA molecule is given by the sequence of its nucleotides, which can be represented as a string over the alphabet ('A', 'C', 'G', 'T') in the case of DNA, and a string over the alphabet ('A', 'C', 'G', 'U') in the case of RNA. Bases with unknown molecular composition are denoted with 'N'.

[SOURCE: ISO 23092-2:2020, 3.20]

3.11
ontology
logical structure of the *terms* (3.20) used to describe a domain of knowledge, including both the definitions of the applicable terms and their relationships

[SOURCE: ISO/IEC/IEEE 24765:2017, 3.2691]

3.12
pedigree
structured description of the familial relationships between samples

Note 1 to entry: Pedigree information is represented in the form of a PED file.

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

3.13
phenopacket
uniform, machine-readable, phenotypic description of an individual, patient or sample

Note 1 to entry: Includes a catch-all collection of data types, specifically focused on representing disease data in both initial data capture and analysis.

3.14
phenotype
set of observable characteristics of an organism resulting from the interaction of its genotype with the environment

Note 1 to entry: 'Phenotypic feature' is a descriptive feature, such as Arachnodactyly, that is the component of a disease, such as Marfan syndrome. It can be observed as either present or absent (excluded), with possible onset, modifiers and frequency.

3.15
proband
affected family member who seeks medical attention thereby bringing the family under study

3.16
quality score
quality value
number assigned to each *nucleotide base* (3.10) call in automated sequencing processes

Note 1 to entry: Quality values express the base-call accuracy, i.e. the probability (or a related measure) for a nucleotide in the sequence to have been incorrectly determined.

[SOURCE: ISO 23092-2:2020, 3.22, modified — First preferred term and admitted term have been swapped.]

3.17
reference sequence
nucleic acid sequence used either to align by mapping sequence reads or as the basis for annotations such as genes and sequence variations

[SOURCE: ISO 20397-2:2021, 3.26]

3.18
single nucleotide polymorphism
SNP
single nucleotide variation in a genetic sequence that occurs at appreciable frequency in the population

Note 1 to entry: Pronounced "snip".

Note 2 to entry: Can also be referred to as single nucleotide variation (SNV).

[SOURCE: ISO 25720:2009, 4.23, modified — Note 1 and Note 2 to entry have been added.]

3.19
string
data type consisting of a sequence of one or more characters

[SOURCE: ISO 2146:2010, 4.6.9]

3.20
term
ontology class composed of a definition, a label, and a unique identifier

3.21
UBERON
comparative anatomy ontology representing a variety of structures found in animals, such as lungs, muscles, bones, feathers and fins

© ISO 2022 – All rights reserved

AVA Documents Document Sharing Hub
Full document details: BS ISO 4454:2022

3.22
uniform resource identifier
URI
string (3.19) of characters that unambiguously identifies a particular resource, such as registered name spaces or protocols

3.23
variant
alteration in the most common DNA nucleotide sequence

Note 1 to entry: It can describe an alternation that can be benign, pathogenic, or of unknown significance.

Note 2 to entry: Variant implies deletion, insertion, indel or single nucleotide polymorphism.

4 Abbreviated terms

ACMG	American College of Medical Genetics
AJCC	American Joint Committee on Cancer
CNV	Copy Number Variation
DNA	Deoxyribonucleic Acid
ECO	Evidence and Conclusion Ontology
EHR	Electronic Health Record
GENO	Genotype Ontology
HGNC	HUGO Gene Nomenclature Committee
HGVS	Human Genome Variation Society
HPO	Human Phenotype Ontology
HTS	High-Throughput Sequencing
HUGO	Human Genome Organization
ICD	International Classification of Diseases
IRI	Internationalized Resource Identifier
ISCN	International System for Human Cytogenomic Nomenclature
MONDO	Mondo Disease Ontology
NCIT	National Cancer Institute Thesaurus
OBO	Open Biological and Biomedical Ontology
OMIM	Online Mendelian Inheritance in Man
PDX-MI	Patient-derived tumor xenograft minimal information standard
PURL	Persistent Uniform Resource Locator
RNA	Ribonucleic Acid
SAM	Sequence Alignment Map

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

SPDI	Sequence Position Deletion Insertion
TNM	Classification of Malignant Tumors
URL	Uniform Resource Locator
VCF	Variant Call Format
VRS	Variation Representation Specification

5 Phenopackets Schema and Requirements

5.1 Phenopacket Schema

The phenopacket schema contains a common, limited set of data types which can be composed into more specialized types for data sharing between resources using an agreed upon common schema. There are three top-level elements – Phenopacket, Family, and Cohort – with other properties, or 'building blocks', nested within. An overview of schema elements and their thematic groupings can be found in Figure 1, with detailed class diagrams of those thematic groupings shown in Figure 2.

The phenopacket is formally defined in protobuf3[1]. Protobuf is language-neutral, faster than other schema languages such as XML and JSON and can be simpler to use because of features such as automatic validation of data objects. It also works with many languages, including Java, GO, C#, C++, JS and Python.[2] See Annex A for several examples that demonstrate how to work with phenopackets in Java and C++.

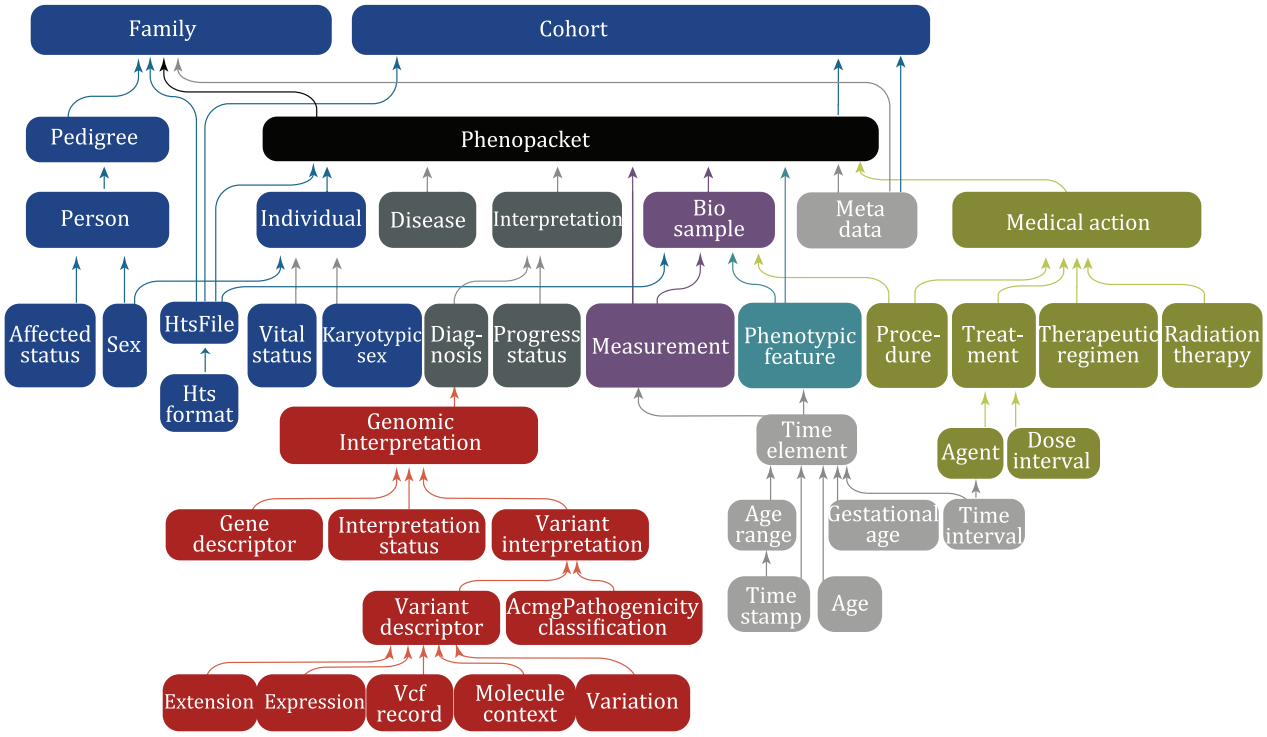
Given the nested nature of phenopackets elements, it can be difficult to understand the overall structure and relationships within phenopackets in a linear document. The documentation for the phenopacket-schema with hyperlinked building blocks can be found at https://phenopacket-schema.readthedocs.io/en/v2/index.html.

1) Protobuf is an exchange format developed by Google LLC. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO.

2) These trademarks are examples of suitable products available commercially. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of these products.

© ISO 2022 – All rights reserved

AVA Documents Document Sharing Hub
Full document details: BS ISO 4454:2022



Key

High-level themes

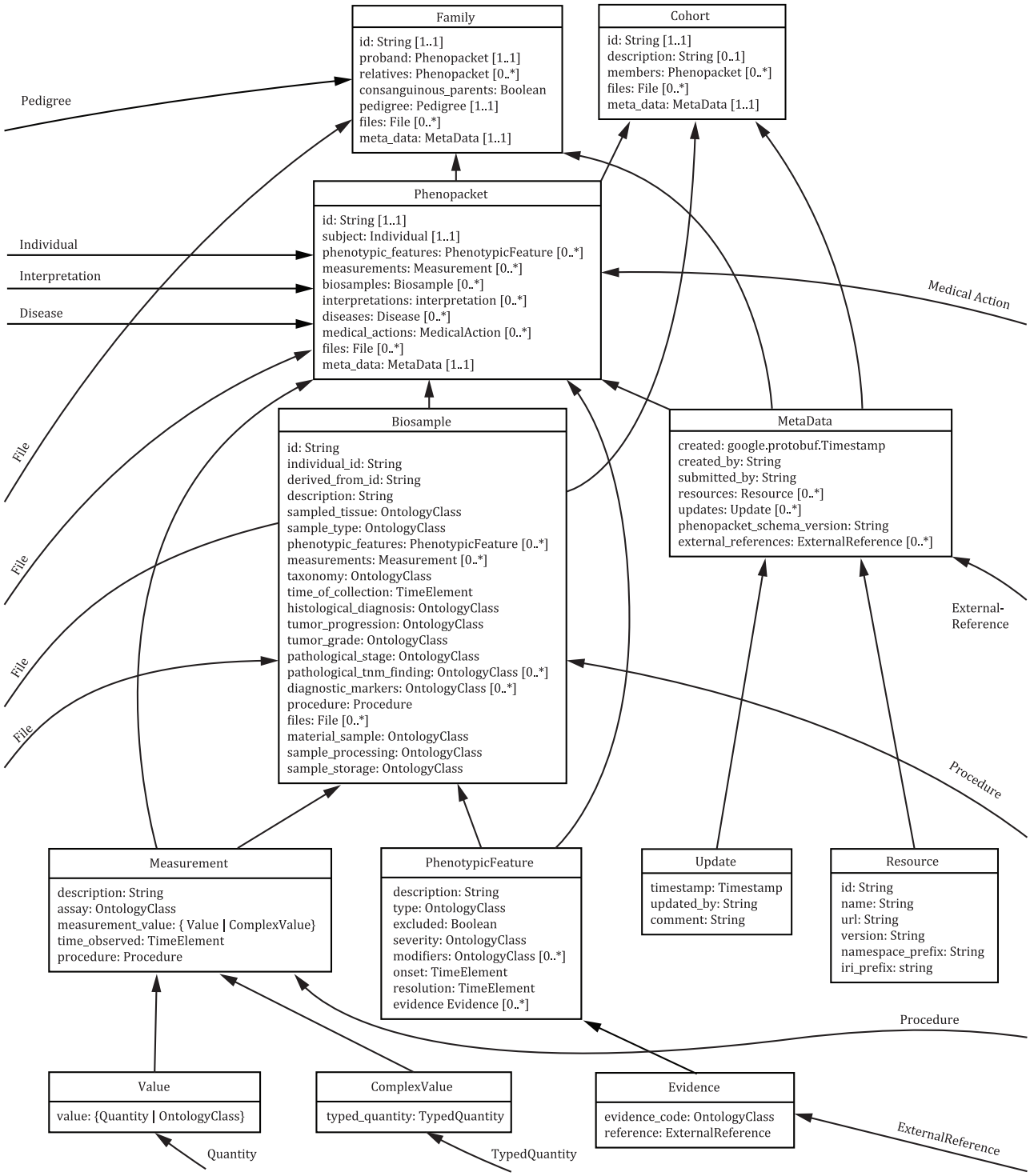
- core phenopacket
- biosample classes
- base classes: person/family/cohort
- disease interpretation classes
- genomic interpretation classes
- medical action classes
- phenotypic features
- multi-purpose classes

Arrows between elements indicate composition. Some classes have been omitted for legibility. See Figure 2 for a detailed class diagram of each thematic grouping.

Figure 1 — Simplified Overview of Phenopackets Schema Elements

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

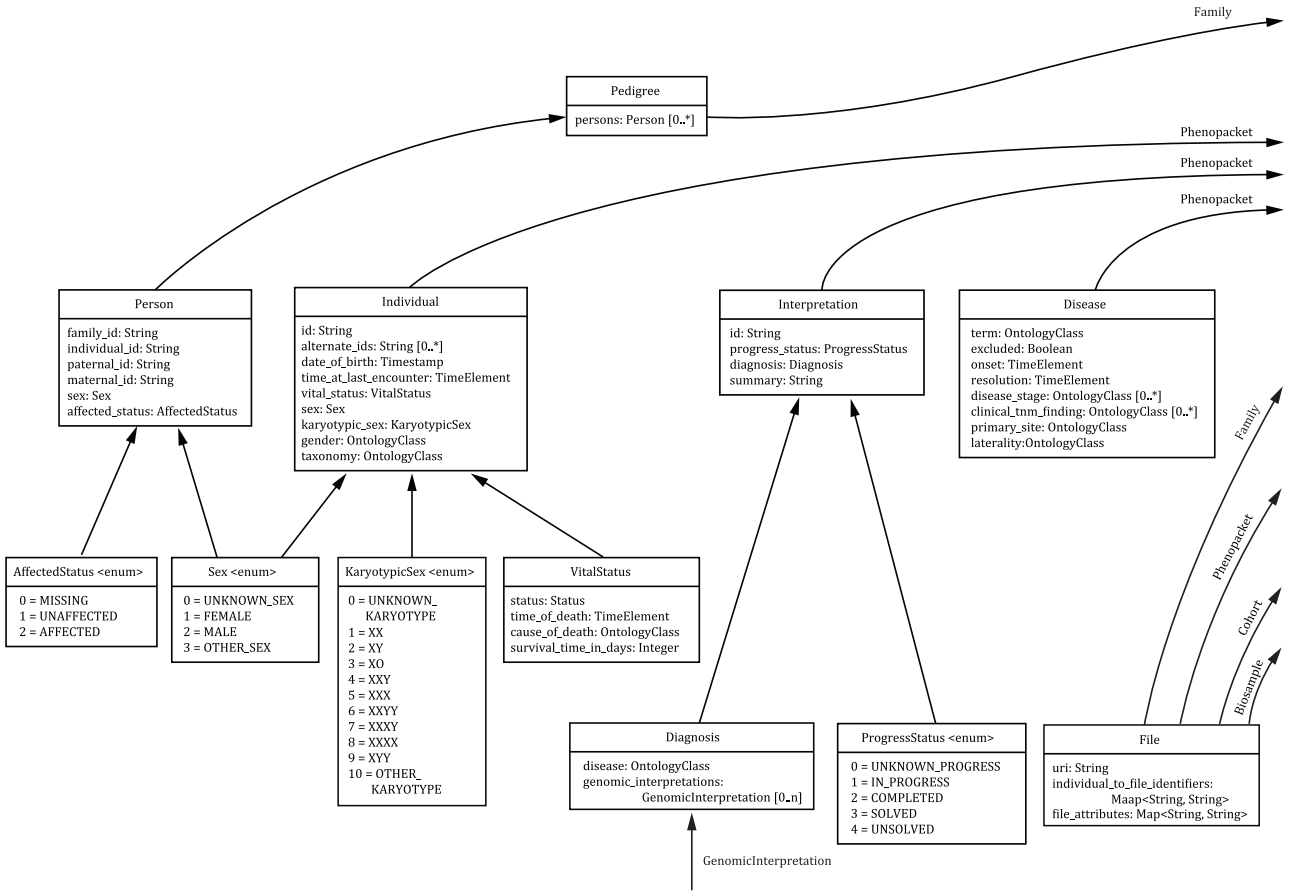


Arrows between elements indicate composition. The OntologyClass and TimeElement links have been omitted for legibility.

a) Detailed view of the top-level base classes of phenopackets, biosample classes and measurement classes

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

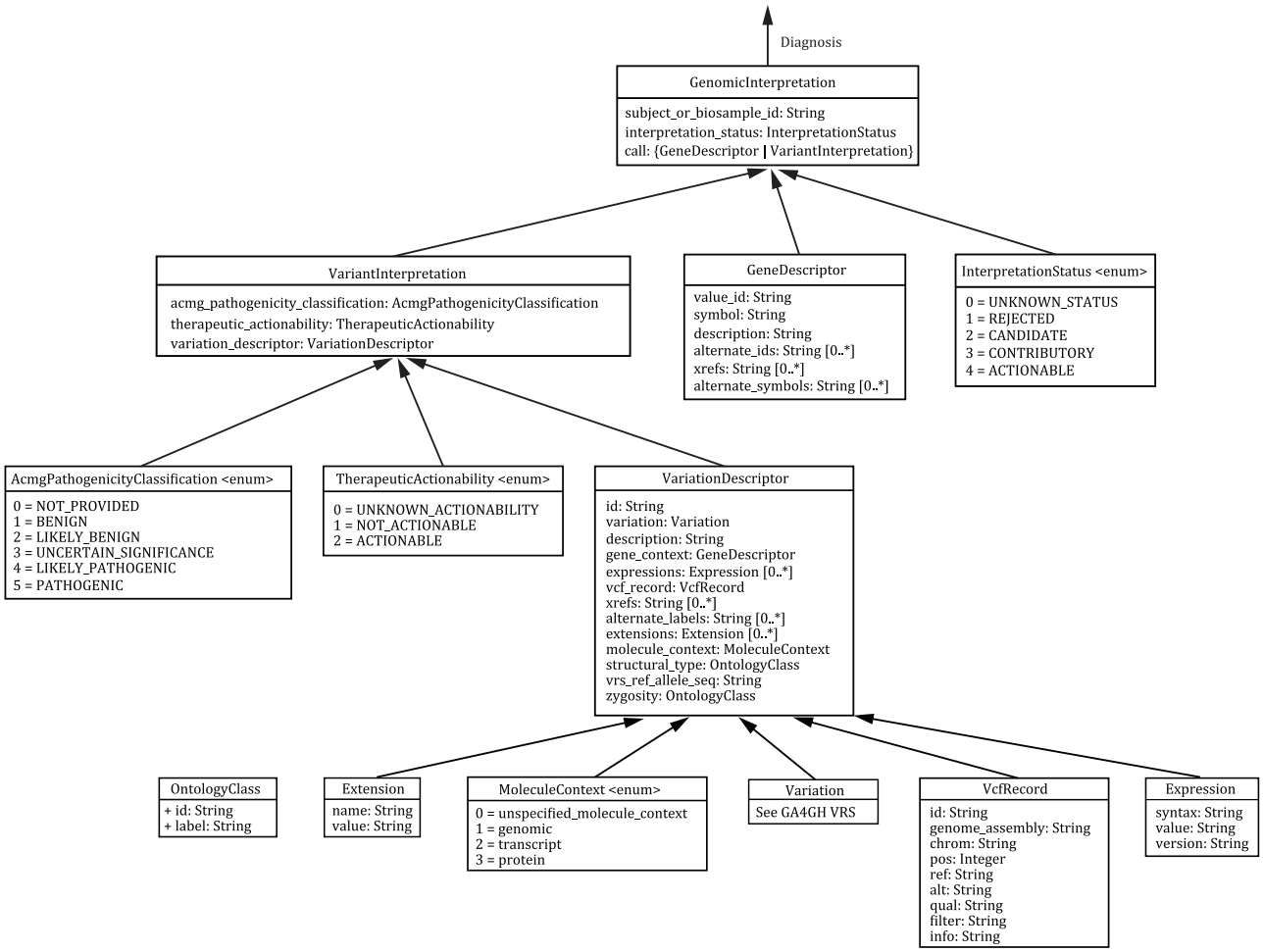


Arrows between elements indicate composition. The OntologyClass and TimeElement links have been omitted for legibility.

b) Detailed view of the base classes and disease/interpretation classes of the phenopackets schema

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

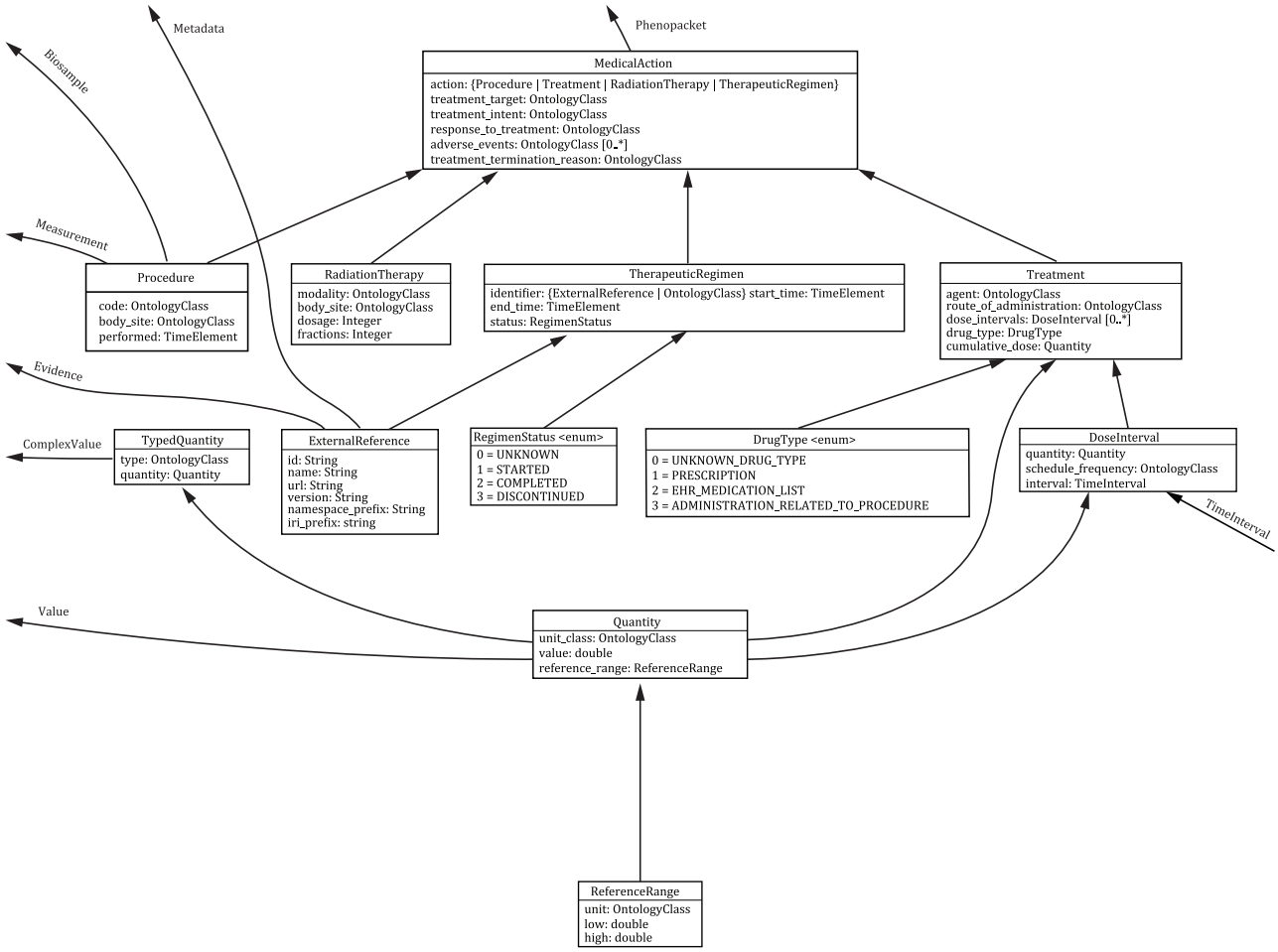


Arrows between elements indicate composition. The OntologyClass links have been omitted for legibility.

c) Detailed view of the genomic interpretation classes of the phenopackets schema

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

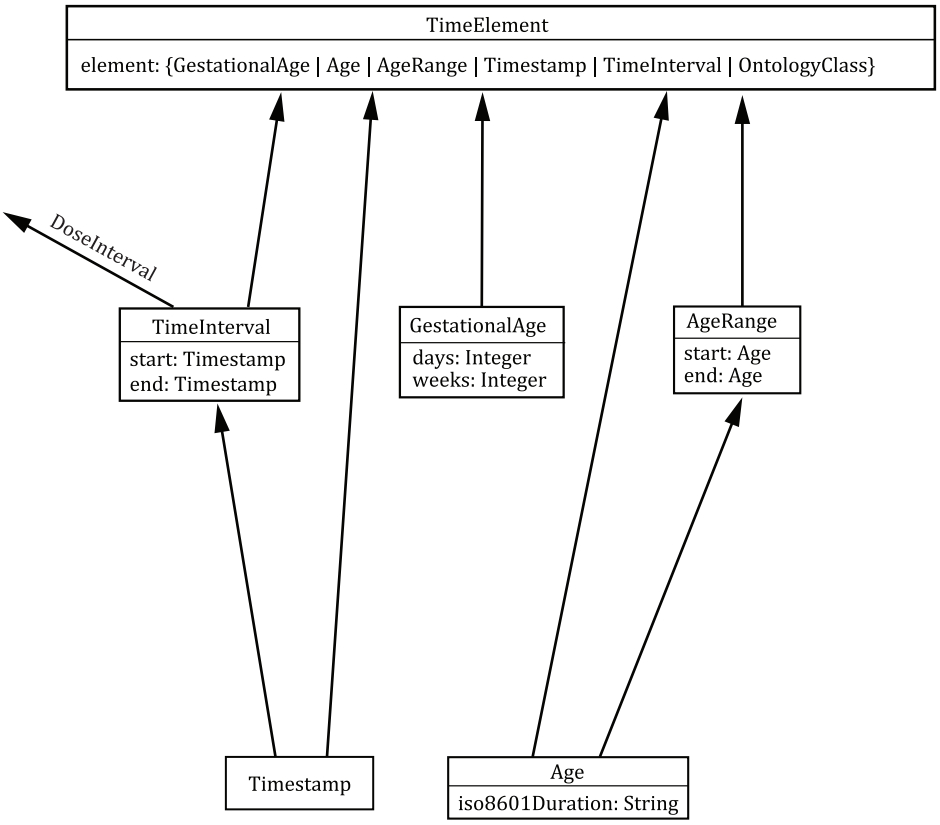


Arrows between elements indicate composition. The OntologyClass and some TimeElement links have been omitted for legibility.

d) Detailed view of the medical action classes of the phenopackets schema

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022



Arrows between elements indicate composition. The OntologyClass links have been omitted for legibility.

e) Detailed view of the multi-purpose time classes of the phenopackets schema

Figure 2 — Phenopackets Schema Class Diagram

5.2 Requirement Levels

5.2.1 General

In protobuf3, as all elements are optional, there is no mechanism within protobuf to declare that a certain field is required. The phenopacket schema does require some fields to be present and, in some cases, additionally requires that these fields have a certain format (syntax) or intended meaning (semantics). Software that uses phenopackets should check the validity of the data with other means.

The requirement levels that are shown for the various elements of the phenopacket only apply if the element is used. For instance, the quantity building block shows that the unit and value fields are required (the multiplicity is exactly 1). In contrast, the field reference_range is optional (the multiplicity can be 0 or 1). The requirements only apply if a quantity element is used in a phenopacket. For instance, phenopackets that do not contain measurement or treatment elements do not contain quantity elements, and so the requirements for the fields in a quantity element do not apply.

5.2.2 Multiplicity

Each building block model contains a multiplicity column. The multiplicity value definitions are as follows:

0..1: The element can be absent (0) or present (1). The element is optional.

1..1: The element shall be present (1). The element shall be used.

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

0..*: There can be from zero to an arbitrary number of elements, i.e. a potentially empty list. This element is optional.

1..*: There can be from one to an arbitrary number of elements, i.e. a list that shall not be empty. This element shall be used.

5.3 Ontology Use

The phenopacket schema shall use a common ontology that allows sophisticated algorithmic analysis over medically relevant abnormalities. Multiple ontologies can be used within a phenopacket. See Annex B for example ontologies and terms (see 3.21).

EXAMPLE HPO, NCIT, SNOMED

6 Phenopacket Schema Top-Level Elements

6.1 Phenopacket

6.1.1 General

The Phenopacket top-level element can be used to describe the phenotypic characteristics observed in an individual with a disease that is being studied or for an individual in whom the diagnosis is being sought (see Table 1). A phenopacket can contain information about genetic findings that are causative of the disease, or alternatively it can contain a reference to a VCF file if whole exome sequencing is being performed as a part of the differential diagnostic process. A phenopacket can also be used to describe the constitutional phenotypic findings of an individual with cancer. See Annex C, D, and E for example phenopackets generated for rare disease, cancer, and COVID-19 use cases, respectively.

Table 1 — Phenopacket elements

Field	Type	Multiplicity	Description
id	string	1..1	Arbitrary identifier
subject	Individual	0..1	The proband
phenotypic_features	PhenotypicFeature (list)	0..*	Phenotypic features observed in the proband
measurements	Measurement (list)	0..*	Measurements performed on the proband
biosamples	Biosample	0..*	Samples (e.g. biopsies), if any
interpretations	Interpretation (list)	0..*	Interpretations related to the phenopacket
diseases	Disease (list)	0..*	Disease(s) diagnosed in the proband
medical_actions	MedicalAction (list)	0..*	Medical actions performed on the proband
files	File (list)	0..*	A list of files related to the subject
meta_data	MetaData	1..1	Information about ontologies and references used in the phenopacket

6.1.2 id

The id is an identifier specific for this phenopacket. The syntax of the identifier is application specific.

6.1.3 subject

This is the individual human (or another organism) that the phenopacket is describing. In many cases, the individual will be a patient or proband of the study. See 7.15 (Individual) for further information.

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

6.1.4 phenotypic_features

This is a list of phenotypic findings observed in the subject. See 7.23 (PhenotypicFeature) for further information.

6.1.5 measurements

A list of measurements performed in the patient. In contrast to 7.23 (PhenotypicFeature), which relies on 7.21 (OntologyClass) to specify the observation, the measurement element can be used to report quantitative as well as ordinal or categorical measurements. See 7.18 (Measurement) for further information.

6.1.6 biosamples

This field describes samples that have been derived from the patient who is the object of the phenopacket, or a collection of biosamples in isolation. See 7.4 (Biosample) for further information.

6.1.7 interpretation

An optional list of interpretations related to the phenopacket. This element is intended to represent interpretation of disease or phenotypic findings based on genomic findings and shall relate to either a genetic or genomic investigation of organismal origin (e.g. germline DNA derived from a blood sample) or from a biosample. See 7.16 (Interpretation) for further information.

6.1.8 diseases

This is a field for disease identifiers and can be used for listing either diagnosed or suspected conditions. The resources using these fields should define what this represents in their context. See 7.6 (Disease) for further information.

6.1.9 medical_actions

A list of treatments or other medical actions performed for proband or individual represented by the phenopacket. See 7.19 (MedicalAction) for further information.

6.1.10 files

This element contains a list of pointers to the relevant file(s) for the subject. For example, this could include VCF or the results of another high-throughput sequencing experiment. See 7.11 (File) for further information.

6.1.11 meta_data

This element contains structured definitions of the resources and ontologies used within the Phenopacket. It is expected that every valid Phenopacket contains a MetaData element. See 7.20 (MetaData) for further information.

6.2 Family

6.2.1 General

The optional Family element is used to represent phenotype, sample and pedigree data required for a genomic diagnosis (see Table 2). In many cases, genetic diagnostics of Mendelian and other disease is performed on a family basis in order to check for co-segregation of candidate variants with a disease. Usually, one family member is designated as the proband, for instance, a child affected by a genetic disease can be the proband in a family. Genomic diagnostics can involve genome sequencing of the child and his or her parents. In this case, the Family element would include a Phenopacket for the child

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

(proband field). If the parents themselves display phenotypic findings relevant to the analysis, then Phenopacket elements are included for them (using the relatives field). If the parents do not display any relevant phenotypic findings, then it is not necessary to include Phenopacket elements for them. Instead, their status as unaffected can be recorded with the Pedigree element.

Table 2 — Family elements

Field	Type	Multiplicity	Description
id	string	1..1	Application-specific identifier
proband	Phenopacket	1..1	The proband (index patient) in the family
relatives	Phenopacket (list)	0..*	List of phenopackets for family members other than the proband
pedigree	Pedigree	1..1	Representation of the pedigree
files	File (list)	0..*	A list of files related to the whole family, e.g. multi-sample high-throughput sequencing files
meta_data	MetaData	1..1	Metadata about ontologies and references used in this message

6.2.2 id

An identifier specific for this family.

6.2.3 proband

The individual representing the focus of this packet. See 6.1 (Phenopacket) for further information.

6.2.4 relatives

Individuals related in some way to the patient. The individuals can be genetically related or can be members of a cohort. If this field is used, then it is expected that a pedigree will be included for genetically related individuals for use cases such as genomic diagnostics. If a phenopacket is being used to describe one member of a cohort, then in general one Phenopacket will be created for each of the individuals in the cohort. If all that is required is to record affected/not affected status in a family, it is possible to use the Pedigree element only.

6.2.5 pedigree

The pedigree defining the relations between the proband and their relatives. This element contains information compatible with the information in a PED file. The individual_id in the Pedigree element shall map to the id in the Phenopacket element. See 7.22 (Pedigree) for further information.

6.2.6 files

This element contains a list of pointers to the relevant file(s) for the family as a whole. The file(s) shall refer to the entire family. Otherwise, individual files shall be contained within their appropriate scope, for example, within a Phenopacket for germline samples of an individual or within the scope of the Biosample element in the case of data derived from that biosample.

In the case of multi-sample high-throughput sequencing files the sample identifiers in the high-throughput sequencing file shall each map to individual_id referenced in the Pedigree element, in order that linkage analysis can be performed on the sample. See 7.11 (File) for further information.

6.2.7 meta_data

This element contains structured definitions of the resources and ontologies used within the phenopacket. It is expected that every valid Phenopacket contains a MetaData element. See 7.20 (MetaData) for further information.

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

6.3 Cohort

6.3.1 General

This element describes a group of individuals related in some phenotypic or genotypic aspect. For instance, a cohort may consist of individuals all of whom have been diagnosed with a certain disease or have been found to have a certain phenotypic feature. See Table 3 for the fields in the Cohort top-level element.

It is recommended to use the Family element to describe families being investigated for the presence of a Mendelian disease.

Table 3 — Cohort elements

Field	Type	Multiplicity	Description
id	string	1..1	Arbitrary identifier for the cohort
description	string	0..1	Arbitrary text description of the cohort
members	Phenopacket (list)	1..*	Phenopackets that represent members of the cohort
files	File (list)	0..*	A list of files related to the whole cohort (e.g. multi-sample high-throughput sequencing files)
meta_data	MetaData	1..1	Metadata related to the ontologies and references used in this message

6.3.2 id

The id is an identifier specific for this cohort. The syntax of the identifier is application specific.

6.3.3 description

Any information relevant to the study or cohort may be added here as free text.

6.3.4 members

One Phenopacket is included for each member of the cohort. See 6.1 (Phenopacket) for more information.

6.3.5 files

This element contains a list of pointers to the relevant HTS file(s) for the cohort. The file(s) shall be a multi-sample file referring to the entire cohort, if appropriate. Otherwise, individual files shall be contained within their appropriate scope, for example, within a Phenopacket for germline samples of an individual or within the scope of a Biosample in the case of data derived from sequencing that biosample. See 7.11 (File) for further information.

6.3.6 meta_data

This element contains structured definitions of the resources and ontologies used within the phenopacket. It is expected that every valid Phenopacket contains a MetaData element. See 7.20 (MetaData) for further information.

7 Phenopackets Building Blocks

7.1 General

The phenopacket standard consists of several protobuf messages each of which contains information about a certain topic such as phenotype, variant, and pedigree. One message can contain other messages, which allows a rich representation of data. For instance, the phenopacket message contains

© ISO 2022 – All rights reserved

AVA Documents Document Sharing Hub
Full document details: BS ISO 4454:2022

messages of type Individual, PhenotypicFeature, Biosample, and so on. Individual messages can therefore be regarded as building blocks that are combined to create larger structures. It would also be straightforward to include the phenopackets schema into larger schema for particular use cases.

7.2 Age

The Age element allows the age of the subject to be encoded in several different ways that support different use cases. Age is encoded as ISO 8601 duration, in accordance with the ISO 8601 series. See Table 4 for the fields in an Age element.

Table 4 — Age data model

Field	Type	Multiplicity	Description
iso8601duration	string	1..1	An ISO 8601 string representing age

If the Age message is used, the iso8601duration value shall be present. The string element (string age=1) should be used for ISO 8601 durations (e.g. P40Y10M05D). For many use cases, less precise strings will be preferred for privacy reasons, e.g. P40Y.

EXAMPLE

```
age:
  iso8601duration: "P25Y3M2D"
```

7.3 AgeRange

The AgeRange element is intended to be used when the age of a subject is represented by a bin, e.g. 5 years old to 10 years old. Bins such as this are used in some situations in order to protect the privacy of study participants, whose age is then represented by bins such as 45 years old to 49 years old, 50 years old to 54 years old, 55 years old to 59 years old, and so on. See Table 5 for the fields in an AgeRange element.

Table 5 — AgeRange data model

Field	Type	Multiplicity	Description
start	Age	1..1	An Age message
end	Age	1..1	An Age message

EXAMPLE To represent the bin 45 years old to 49 years old, one can use an Age element with **P45Y** as the start element of the AgeRange element, and an Age element with **P49Y** as the end element. An AgeRange.end shall occur after AgeRange.start.

```
ageRange:
  start:
      iso8601duration: "P45Y"
  end:
      iso8601duration: "P49Y"
```

7.4 Biosample

7.4.1 General

This element represents a biosample (see Table 6). Examples would be a tissue biopsy, a single cell from a culture for single cell genome sequencing or a protein fraction from a gradient centrifugation. Several instances (e.g. technical replicates) or types of experiments (e.g. genomic array as well as RNA-seq experiments) can refer to the same biosample.

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

Table 6 — Biosample data model

Field	Type	Multiplicity	Description
id	string	1..1	Arbitrary identifier
individual_id	string	0..1	Arbitrary identifier
derived_from_id	string	0..1	Id of the biosample from which the current biosample was derived (if applicable)
description	string	0..1	Arbitrary text
sampled_tissue	OntologyClass	0..1	Tissue from which the sample was taken
sample_type	OntologyClass	0..1	The type of biosample (e.g. RNA, DNA, Cultured cells)
phenotypic_features	PhenotypicFeature (list)	0..*	List of phenotypic abnormalities of the sample
measurements	Measurement (list)	0..*	List of measurements related to the sample
taxonomy	OntologyClass	0..1	Species of the sampled individual
time_of_collection	Time Element	0..1	Age of the proband at the time the sample was taken
histological_diagnosis	OntologyClass	0..1	Disease diagnosis that was inferred from the histological examination
tumor_progression	OntologyClass	0..1	Indicates primary, metastatic, recurrent
tumor_grade	OntologyClass (list)	0..1	List of terms to indicate the grade of the tumor
pathological_stage	OntologyClass	0..1	Pathological stage of the tumor, if applicable
Pathological_tnm_finding	OntologyClass (list)	0..*	Pathological TNM findings, if applicable
diagnostic_markers	OntologyClass	0..*	Clinically relevant biomarkers
procedure	Procedure	0..1	The procedure used to extract the biosample
hts_files	HtsFile (list)	0..*	List of high-throughput sequencing files derived from the biosample
material_sample	OntologyClass	0..1	Status of the specimen (normal control, tumor tissue, etc.)
sample_processing	OntologyClass	0..1	How the specimen was processed
sample_storage	OntologyClass	0..1	How the specimen was stored

EXAMPLE The staging system most often used for bladder cancer is the AJCC TNM system. The overall stage is assigned based on the T, N, and M categories (Cancer stage grouping). For instance, stage II (pathological staging) is defined as T2a or T2b, N0, and M0, meaning the cancer has spread into the wall of the bladder.

```
biosample:
  id: "sample1"
  individualId: "patient1"
  description: "Additional information can go here"
  sampledTissue:
      id: "UBERON_0001256"
      label: "wall of urinary bladder"
  histologicalDiagnosis:
      id: "NCIT:C39853"
      label: "Infiltrating Urothelial Carcinoma"
  tumorProgression:
      id: "NCIT:C84509"
      label: "Primary Malignant Neoplasm"
  tumorGrade:
      id: "NCIT:C36136"
      label: "Grade 2 Lesion"
  procedure:
      code:
          id: "NCIT:C5189"
          label: "Radical Cystoprostatectomy"
  files:
      - uri: "file:///data/genomes/urothelial_ca_wgs.vcf.gz"
```

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

```
    individualToFileIdentifiers:
        patient1: "NA12345"
    fileAttributes:
        description: "Urothelial carcinoma sample"
        htsFormat: "VCF"
        genomeAssembly: "GRCh38"
materialSample:
    id: "EFO:0009655"
    label: "abnormal sample"
timeOfCollection:
    age:
        iso8601duration: "P52Y2M"
pathologicalStage:
    id: "NCIT:C28054"
    label: "Stage II"
pathologicalTnmFinding:
- id: "NCIT:C48726"
    label: "T2b Stage Finding"
- id: "NCIT:C48705"
    label: "N0 Stage Finding"
- id: "NCIT:C48699"
    label: "M0 Stage Finding"
```


7.4.2 id

The biosample ID. This is unique in the context of the server instance.

7.4.3 individual_id

The ID of the Individual this biosample was derived from. It is recommended to provide this information here if the Biosample element is being transmitted as a part of a phenopacket.

7.4.4 derived_from_id

The ID of the parent biosample this biosample was derived from, if applicable.

7.4.5 description

The biosample's description. This attribute contains human readable text. The "description" attributes should not contain any structured data.

7.4.6 sampled_tissue

An OntologyClass describing the tissue from which the specimen was collected. UBERON should be used. The PDX-MI mapping is Specimen tumor tissue.

7.4.7 sample_type

An OntologyClass describing the type of same sample (e.g. RNA, DNA, or Cultured cells). It is recommended to use an EFO term to describe the sample, for instance, genomic DNA (EFO:0008479).

7.4.8 phenotypic_features

The phenotypic characteristics of the biosample, for example histological findings of a biopsy. See 7.23 (PhenotypicFeature) for further information.

7.4.9 measurements

Measurements (usually quantitative) performed on the sample. See 7.18 (Measurement) for further information.

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 4454:2022

7.4.10 taxonomy

An OntologyClass representing the taxonomic identifier of the organism. For resources where there is more than one organism being studied it is advisable to indicate the taxonomic identifier to its most specific level. It is recommended to use codes from the NCBI Taxonomy resource.

EXAMPLE NCBITaxon:9606 (homo sapiens), NCBITaxon:9615 (dog).

7.4.11 time_at_collection

An age object describing the age of the individual this biosample was derived from at the time of collection. The Age object allows the encoding of the age either as ISO 8601 duration or time interval (preferred), or as ontology term object. See 7.31 (TimeElement) for further information.

7.4.12 histological_diagnosis

This is the pathologist's diagnosis and can represent a refinement of the clinical diagnosis (which could be reported in the Phenopacket that contains this biosample). Normal samples would be tagged with the term "NCIT:C38757", "Negative Finding". See 7.21 (OntologyClass) for further information.

7.4.13 tumor_progression

This field can be used to indicate if a specimen is from the primary tumor, a metastasis or a recurrence. There are multiple ways of representing this using ontology terms, and the terms chosen should have a specific meaning that is application specific.

EXAMPLE A term from the following NCIT terms from the Neoplasm by Special Category can be chosen: Primary Neoplasm, Metastatic Neoplasm or Recurrent Neoplasm

7.4.14 tumor_grade

This should be a child term of NCIT:C28076 (Disease Grade Qualifier) or equivalent.

7.4.15 pathological_stage

An OntologyClass to represent the pathological stage of the sample, if applicable.

7.4.16 pathological_tnm_finding

A list of OntologyClass to represent the pathological TNM findings, if applicable. The TNM staging system is used to describe the amount and spread of cancer in a patient's body.

7.4.17 diagnostic_markers

A field for clinically relevant bio markers. Most of the assays, such as immunohistochemistry, are covered by the NCIT under the sub-hierarchy NCIT:C25294 (Laboratory Procedure).

EXAMPLE NCIT:C68748 (HER2/Neu Positive), NCIT:C131711 (Human Papillomavirus-18 Positive).

7.4.18 procedure

The clinical procedure performed on the subject in order to extract the biosample. See Procedure (7.24) for further information.

7.4.19 files

This element contains a list of pointers to relevant file(s) for the biosample. For example, this could include the results of a high-throughput sequencing experiment. See 7.11 (File) for further information.

Documents
Document Sharing Hub

© ISO 2022 – All rights reserved

Full document details: BS ISO 4454:2022