
**Genomics informatics —
Phenopackets: A format for
phenotypic data exchange**

iTeh STANDARD PREVIEW
(standards.iteh.ai)

ISO 4454:2022

<https://standards.iteh.ai/catalog/standards/sist/ad6a117b-0a90-4ef4-a1fa-bfcab8280266/iso-4454-2022>



iTeh STANDARD PREVIEW
(standards.iteh.ai)

ISO 4454:2022

<https://standards.iteh.ai/catalog/standards/sist/ad6a117b-0a90-4ef4-a1fa-bfcab8280266/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

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

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

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

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

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

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*.

ISO 4454:2022

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.

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).

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)

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.

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

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

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 protobuf³¹⁾. 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>.

<https://standards.iteh.ai/catalog/standards/sist/ad6a117b-0a90-4ef4-a1fa-bfcab8280266/iso-4454-2022>

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.