INTERNATIONAL STANDARD

ISO 4454

First edition
2022-07

Genomics informatics — Phenopackets: A format for phenotypic data exchange

Reference number
ISO 4454:2022(E)

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

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.

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 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.