ISO/DTS 8392~~-2022~~.2:2023 (E)

Date: ~~2022-08-15~~2023-01-05

ISO/TC 215/SC 1

Secretariat: KATS

Genomics informatics — Description rules for genomic data for genetic detection products and services

ISO copyright office

CP 401 • Ch. de Blandonnet 8

CH-1214 Vernier, Geneva

Phone: +41 22 749 01 11

Fax: +41 22 749 09 47

Email: copyright@iso.org

Website: www.iso.org

Published in Switzerland

Contents

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

The decreasing cost of sequencing and the gradual in-depth study of genomics have led to the generation of more and more genomic data, but the data quality in genomics is not optimal. From the dimension of data level, there is a lack of data integrity, and medical information has been facing a problem of semantic disunity. These problems have caused great obstacles to downstream applications.

Standardization of data is a prerequisite for data asset management and data storage and applications, which can give better storage for genomic data and enlarge these genomic data used in precision medicine.

This document is based on the actual situation of industry data production, combined with the needs of upstream and downstream industry users. It also takes into account the use made by stakeholders and user friendliness for all common types of genomic data. Solving the problem of data scope and semantic unification can enhance the data association ability, ensure information exchange, improve data flow, improve the data quality from the aspects of data integrity and data validity, and lay a good foundation for subsequent data storage, data application and data sharing.

Genomics informatics — Description rules for genomic data for genetic detection products and services

1 Scope

This document specifies requirements on the category definition and quality assessment of genomic data, including the content structure, attribute and description rules of data format, and the compilation rules of data format.

This document applies to all the genomic data used for human genetic detection products and services.

This document applies to genomic data processing and analysis~~.~~, and to the quality evaluation/assessment of genomic data.

2 Normative references

There are no normative references in this document.

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
alignment-sequence code
continuous coding of objects in the same series, and reserving of extended space

3.2
code
representation of a piece of information such as a letter, word or phrase in another form, usually briefer

3.3
code structure
representation of the composition and length of a complete code

3.4
equal length code
coding system in which all coding objects have the same length

3.5
data identifier
DI
identifier that uniquely distinguishes one set of data from all others

3.6
layer code
hierarchical code consisting of membership order of coded objects

3.7
sequential code
code that represents in the natural order of Arabic numerals, or letters

3.8
variable-length code
code system in which the length of code is not exactly the same

3.9
version identifier
VI
unique number assigned to identify a version of submitted genomic data

4 Data format attribute and description rules

Genomic data can be classified as unstructured data and structured data.

The unstructured data should be described by data format, illustration of data format and archiving catalogue.

The structured data should be described by metadata and data element code.

5 Composition and rules of genomic data description

5.1 Identifier

The description of genomic data should include data format, data attribute and metadata.

5.2 Data format

Data attribute elements for description of data format are totally classified of 1411 attributes in five categories, shown in Table A.1. According to the universal property, there include data element common attribute and data element specific attribute.

5.3 Data archiving catalogue

Data elements for data archiving catalogue are totally classified of 1411 attributes in five categories, shown in Table A.2. According to the universal property, there include data element common attribute and data element specific attribute.

5.4 Metadata

Data elements for metadata description are totally classified of 14 attributes in five categories, shown in Table A.3. According to the universal property, there include data element common attribute and data element specific attribute.

6 Core elements and rules for the description of genomic data

6.1 Identifier

Identifier shall use alphanumeric code with. The structure may be considered as two-level structure, including DI and VI.

EXAMPLE 1 DI_V1

a) DI consists of alphanumeric characters combining classification code and serial number:

— A category code consists of two uppercase letters;

— A group code is a 2-digit code. The numerical value has no meaning;

— A class code is a 2-digit code. The numerical value has no inherent meanings to humans. If there is no class, the class code is 00. The group code and the class code are separated by a dot;

— A sequential code is a 3-digit code. It represents the data element number under a class, starting from 001. The numerical value has no meaning.

2) VI consists of 4 elements. It follows the structure "V"+"m..m"+"."+"n..n", where "m..m" and "n..n" are numbers assigned in increasing order. They should be positive integer. "m..m" represents major version numbers and "n..n" represents minor version numbers.

EXAMPLE 2 "V1.2" means the first major version and the second minor version.

— If a valid data exchange can be performed between the versions before and after the data element is updated, the updated major version number remains unchanged, and the minor version number equals to the current minor version number plus one.

— If a valid data exchange cannot be performed between the versions before and after the data element is updated, the updated major version number equals to the current major version number plus one, and the minor version number is returned to zero.

The structure of a data identifier is shown as Figure 1. Data identifier examples are shown in Annex B, such as sequence information (see Table B.1), bioinformatic analysis (see Table B.2).

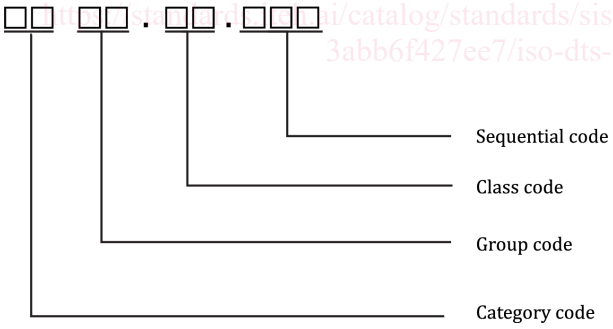

Figure 1 — Structure of data identifier

6.2 Name

6.2.1 The data format name shall be unique and in the form of strings with letters and numbers. The naming of data elements should use a certain logical structure and general terminology.

6.2.2 A complete data element name shall consist of object class term, property term, representation term and (qualifier term).

— A data element has one and only one object class term. If there is one object in an omics data element catalogue, it may be omitted as appropriate;

— A data element has one and only one property term. Property term is an essential component of any data element name. Other terms may be abbreviated as appropriate when the expression of the data element concept is complete, accurate, and unambiguous;

— A data element has a unique representation term. Redundant words can be removed from the name when there are duplicates or partial repetitions of the representation term and the property term;

— Qualifier term is optional and is given from particular professional fields.

7 Requirement of code

7.1 Code structure

7.1.1 The structure design shall follow the requirements:

— The structure of the code shall be concise and avoid carrying too much information;

— The structure shall accord with the basic method of information processing and harmonize with relevant standard structures;

— When adding, deleting or modifying one part of the code, the structure shall be unbroken;

— The code shall use user-friendly symbols.

7.1.2 The description for the data element code structure shall conform with the following requirements:

— The type of code, the structure of the code and the coding method shall be clearly described;

— When the code's structure is too complex, it shall be illustrated using example;

7.1.3 The structure of layer code can also be represented by a schematic diagram, as shown in Figure 2.

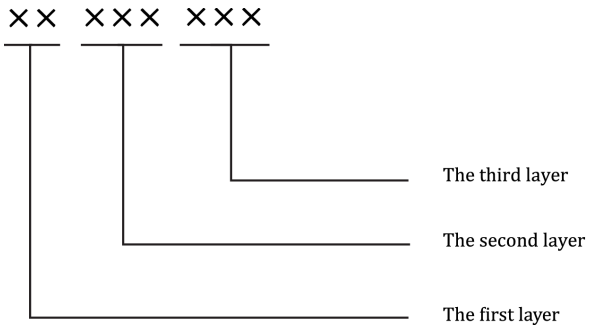

Figure 2 — Diagram of layer code structure

Examples of code lists are given in Annex C.

7.2 Code length

7.2.1 The code length shall meet the demand and be as short as possible. The code should use equal length code but not variable-length code. The headspace of the code should conform with the development of the coding object.

7.2.2 The formula of code length is shown in Annex D.

7.3 Code type and format

7.3.1 Code characters should be numeric, alphabetic or alphanumeric.

7.3.2 Code characters shall be correct and readable. They shall avoid using confusing and misunderstood characters. Characters with similar sounds and shapes in a given code, e.g. “1” and “I”, should be avoided. The code shall be written in the same form, including the uppercase and lowercase letters, font and size

7.3.3 It is recommended to use full numbers or full letters to represent the code. Mixed alphanumeric forms are generally used in special situations or locations. It is not suitable for random use.

7.3.4 Sequential codes should use codes of equal lengths. For example, use 001 to 999 instead of 1 to 999. The code in the same layer should be the same length.

7.3.5 In numeric codes, if there is a host category, they shall use a code with the last number "9".

7.3.6 The code shall be written in the same form, including letter case, font and size.

7.4 Code list naming

7.4.1 The code list shall have an authoritative name in the context of a particular domain.

7.4.2 The name of the code table shall accurately reflect the character in the table as one of the data elements representing class attributes, and shall not enlarge or narrow its scope of use.

7.4.3 The name of the code list shall be concise, convey clear semantics, and reflect the essence of the code list.

8 Compatibility with other rules

Data exchange between different description rules needs data cleaning, especially structured data.

Archiving catalogue should be unanimouschecked and compared for unstructured data with other rules.

Structured data should be complete the missing information, modify the logical error data, remove the unneeded data, and verify the association between metadata and modify inconsistent data element code.

Data cleaning can be semi-automated.