Genomics informatics - Omics Markup Language (OML) (ISO 21393:2021)

EESTI STANDARDI EESSÕNA

NATIONAL FOREWORD

See Eesti standard EVS-EN ISO 21393:2021 sisaldab Euroopa standardi EN ISO 21393:2021 ingliskeelset teksti.	This Estonian standard EVS-EN ISO 21393:2021 consists of the English text of the European standard EN ISO 21393:2021.
Standard on jõustunud sellekohase teate avaldamisega EVS Teatajas.	This standard has been endorsed with a notification published in the official bulletin of the Estonian Centre for Standardisation and Accreditation.
Euroopa standardimisorganisatsioonid on teinud Euroopa standardi rahvuslikele liikmetele kättesaadavaks 11.08.2021.	Date of Availability of the European standard is 11.08.2021.
Standard on kättesaadav Eesti Standardimis- ja Akrediteerimiskeskusest.	The standard is available from the Estonian Centre for Standardisation and Accreditation.

Tagasisidet standardi sisu kohta on võimalik edastada, kasutades EVS-i veebilehel asuvat tagasiside vormi või saates e-kirja meiliaadressile standardiosakond@evs.ee.

ICS 35.240.80

EUROPEAN STANDARD

NORME EUROPÉENNE

EUROPÄISCHE NORM

EN ISO 21393

August 2021

ICS 35.240.80

English Version

Genomics informatics - Omics Markup Language (OML) (ISO 21393:2021)

Informatique génomique - Langage de balisage Omics (OML) (ISO 21393:2021)

Medizinische Informatik - OMICS Auszeichnungssprache (OML) (ISO 21393:2021)

This European Standard was approved by CEN on 29 September 2020.

EUROPEAN COMMITTEE FOR STANDARDIZATION
COMITÉ EUROPÉEN DE NORMALISATION
EUROPÄISCHES KOMITEE FÜR NORMUNG

CEN-CENELEC Management Centre: Rue de la Science 23, B-1040 Brussels

Ref. No. EN ISO 21393:2021 E

European foreword

This document (EN ISO 21393:2021) has been prepared by Technical Committee ISO/TC 215 "Health informatics" in collaboration with Technical Committee CEN/TC 251 “Health informatics” the secretariat of which is held by NEN.

This European Standard shall be given the status of a national standard, either by publication of an identical text or by endorsement, at the latest by February 2022, and conflicting national standards shall be withdrawn at the latest by February 2022.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. CEN shall not be held responsible for identifying any or all such patent rights.

Any feedback and questions on this document should be directed to the users’ national standards body/national committee. A complete listing of these bodies can be found on the CEN websites.

According to the CEN-CENELEC Internal Regulations, the national standards organizations of the following countries are bound to implement this European Standard: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Republic of North Macedonia, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey and the United Kingdom.

Endorsement notice

The text of ISO 21393:2021 has been approved by CEN as EN ISO 21393:2021 without any modification.

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,* in collaboration with the European Committee for Standardization (CEN) Technical Committee CEN/TC 251, *Health informatics,* in accordance with the Agreement on technical cooperation between ISO and CEN (Vienna Agreement).

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

In this post genomic era, the management of health-related data is becoming increasingly important to both omics research and omics-based medicine.[1] Informational approaches to the management of clinical, image and omics data are beginning to have as much worth as basic, bench top research. In the current electronic world, there are multiple different types of data for healthcare as shown in Figure 1. Besides, nowadays there are many kinds of omics data around the world awaiting effective utilization for human health. The development of data format and message standards to support the interchange of clinical omics data is necessary. Omics data includes omics sequence, sequence variation and other expression data, proteomics data, molecular network, etc. As an entry point, this document focuses on the data exchange.

In the present circumstances, omics is expected to be a key to understand human response to external stimuli such as any kinds of alien invasions, therapies, and the environmental interactions.[2] Bacterial infection is an example of alien invasion, and the responses to the infections are different among the individuals. According to the therapy, the side effects to a drug are different among the patients. These responses are also different in various environments. As a result of recent explosive amount of these omics researches, the huge amounts of experimental data have been accumulating in many databases in various types of data formats. These data are waiting to be used in drug discovery, clinical diagnosis, and clinical researches.

The Markup Language is a set of symbols and rules for their use when doing a markup of a document.[3] The first standardized markup language was ISO 8879 onGeneralized Markup Language (SGML)[4] which has strong similarities with troff and nroff text layout languages supplied with Unix systems. Hypertext Markup Language (HTML) is based on SGML.[5] Extensible Markup Language (XML) is a pared-down version of SGML, designed especially for Web documents.[6] XML acts as the basis for Extensible HTML (XHTML)[7] and Wireless Markup Language (WML)[8] and for standardized definitions of system interaction such as Simple Object Access Protocol (SOAP).[9] By contrast, text layout or semantics are often defined in a purely machine-interpretable form, as in most word processor file formats[10].

Markup Language for the biomedical field, based on XML, has been in development for several decades to enhance the exchange data among researchers. Bioinformatic Sequence Markup Language (BSML),[11] Systems Biology Markup Language (SBML),[12] Cell Markup Language (Cell ML),[13] and Neuro Markup Language (Neuro-ML)[14] are examples of markup languages. Polymorphism Mining and Annotation Programs (PolyMAPr)[15] is centric on SNP and tries to achieve mining, annotation, and functional analysis of public database as dbSNP,[16] CGAP,[17] and JSNP[18] through programming. ISO 25720 Genomic Sequence Variation Markup Language (GSVML) is the first standardized ML for clinical genomic sequence variation data exchange.

The purpose of Omics Markup Language (OML) is to provide a standardized data exchange format for omics in human health.

The recent expansion in omics research has produced large quantities of data held in many databases with different formats. Standardization of data exchange is necessary for managing, analysing and utilizing these data. Considering that omics, especially transcriptomics, proteomics, signalomics and metabolomics, has significant meaning in molecular-based medicine and pharmacogenomics, the data exchange format is key to enhancing omics-based clinical research and omics-based medicine.

Recently, informational approaches have become more important to both omics research and omics-based medicine. The management of omics data is as critical as basic research data in this new era. There are many kinds of omics data around the world, and the time has come to effectively use this omics data for human health. To use this data effectively and efficiently, standards should be developed to permit the interoperable interchange of omics data globally. These standards should define the data format as well as the messages that would be used to interchange and share this data globally.

OML is a base frame of all kinds of clinical omics data. Each omics category will be introduced as a specific add on component part. As an instance, Whole Genome sequence Markup Language will be

a specific add on component part for whole genome sequence data, and Genomic Sequence Variation Markup Language will be a specific add on component part for genomic sequence variation data.

To utilize the internationally accumulated omics data, standards for the interchange of omics data should be defined. These standards should define a data format and exchange messages. Markup Language is a reasonable choice to address this need. As for omics data message handling, Health Level Seven®[1)] Clinical Genomics Work Group[19] has summarized clinical use cases for general omics data. The OML project has contributed to these efforts. Additionally, this work incorporated use cases based on the Japanese millennium project.[20] Based on these contexts and investigations, this document elucidates the needs and the requirements for OML and after then proposes the specification of OML for the international standardization based on the elucidated needs and the requirements.

1) Health Level Seven (HL7) is the registered trademark of Health Level Seven International. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named.

Genomics informatics — Omics Markup Language (OML)

1 Scope

This document is applicable to the data exchange format that is designed to facilitate exchanging omics data around the world without forcing changes of any database schema.

This document specifies the characteristics of OML from the following perspectives.

From an informatics perspective, OML defines the data exchange format based on XML. This document gives guidelines for the specifications of the data exchange format, but this document excludes the database schema itself.

From a molecular side of view, this document is applicable to all kinds of omics data, while this document excludes the details of the molecules (e.g., details of genomic sequence variations or whole genomic sequence). This document is also applicable to the molecular annotations including clinical concerns and relations with other omics concerns.

From an application side of view, this document is applicable to the clinical field including clinical practice, preventive medicine, translational research, and clinical research including drug discovery. This document does not apply to basic research and other scientific fields.

From a biological species side of view, this document is applicable to the human health-associated species as human, preclinical animals, and cell lines. This document does not apply to the other biological species.

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 terminological databases for use in standardization at the following addresses:

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

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

3.1
actor
something or someone who supplies a stimulus to the system

Note 1 to entry: Actors include both humans and other quasi-autonomous things, such as machines, computer tasks and systems.

[SOURCE: ISO 25720:2009, 4.1]

3.2
allele
gene that is found in one of two or more different forms in the same position in a chromosome