SLOVENSKI STANDARD
oSIST prEN ISO 21393:2019

01-september-2019

Zdravstvena informatika - Označevalski jezik OMICS (OML) (ISO/DIS 21393:2019)

Health informatics - Omics Markup Language (OML) (ISO/DIS 21393:2019)

Medizinische Informatik - OMICS Auszeichnungssprache (OML) (ISO/DIS 21393:2019)

Informatique de santé — Langage de balisage Omics (OML) (ISO/DIS 21393:2019)

Ta slovenski standard je istoveten z: prEN ISO 21393

ICS:

35.240.80	Uporabniške rešitve IT v zdravstveni tehniki	IT applications in health care technology

oSIST prEN ISO 21393:2019 **en,fr,de**

DRAFT INTERNATIONAL STANDARD

ISO/DIS 21393

ISO/TC **215** Secretariat: **ANSI**

Voting begins on:
2019-07-16

Voting terminates on:
2019-10-08

Health informatics — Omics Markup Language (OML)

Informatique de santé — Langage de balisage Omics (OML)

ICS: 35.240.80

THIS DOCUMENT IS A DRAFT CIRCULATED FOR COMMENT AND APPROVAL. IT IS THEREFORE SUBJECT TO CHANGE AND MAY NOT BE REFERRED TO AS AN INTERNATIONAL STANDARD UNTIL PUBLISHED AS SUCH.

IN ADDITION TO THEIR EVALUATION AS BEING ACCEPTABLE FOR INDUSTRIAL, TECHNOLOGICAL, COMMERCIAL AND USER PURPOSES, DRAFT INTERNATIONAL STANDARDS MAY ON OCCASION HAVE TO BE CONSIDERED IN THE LIGHT OF THEIR POTENTIAL TO BECOME STANDARDS TO WHICH REFERENCE MAY BE MADE IN NATIONAL REGULATIONS.

RECIPIENTS OF THIS DRAFT ARE INVITED TO SUBMIT, WITH THEIR COMMENTS, NOTIFICATION OF ANY RELEVANT PATENT RIGHTS OF WHICH THEY ARE AWARE AND TO PROVIDE SUPPORTING DOCUMENTATION.

This document is circulated as received from the committee secretariat.

ISO/CEN PARALLEL PROCESSING

Reference number
ISO/DIS 21393:2019(E)

COPYRIGHT PROTECTED DOCUMENT

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

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 125, *Health Informatics*, Subcommittee SC 1, Clinical Genomics.

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 next generation 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. Nowadays there are many kinds of omics data around the world awaiting effective utilization for human health. The hurdle that must be overcome to achieve this goal is the development of data format and message standards to support the interchange of clinical omics data. Omics data includes omics sequence, sequence variation and other expression data, proteomics data, molecular network, etc. As an entry point, this standard 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 Standard Generalized 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 must be developed to permit the interoperable interchange of omics data globally. These standards must define the data format as well as the messages to be used to interchange and share this data globally. This standard addresses those requirements, using a markup language.

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 accumulated omics data among many facilities around the world, standards for the interchange of omics data must be defined. The required standards include defining a data format and exchange messages. Markup Language is the reasonable choice to address this need. As for omics data message handling, Health Level Seven 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 then proposes the specification of OML for the international standardization.

A list of references related this part of ISO/DIS 21393 is given in the bibliography.

Health informatics — Omics Markup Language

1 Scope

OML is a data exchange format designed to facilitate exchanging omics data around the world without forcing changes to existing databases.

From an informatics perspective, OML is an XML-based data exchange format. The data exchange format (e.g., XML schema and DTD) is in scope. The structure of the systems and databases sending or receiving the information schemas are out of the scope.

From a biological perspective, all kinds of omics are in scope, but the details (e.g., details of genomic sequence variations or whole genomic sequence) are out of the scope. Annotations including clinical concerns and relations with other omics concerns are in scope.

The application focus is human health including clinical practice, preventive medicine, translational research, and clinical research including drug discovery. The scope includes health-associated species, including human and preclinical animals, and associated cell lines. Other species, basic research, and other scientific fields are out of scope.

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 25720:2009, *Health informatics -- Genomic Sequence Variation Markup Language (GSVML)*

ISO/HL7 21731:2006, *Health informatics – HL7 version 3 – Reference information model – Release 1*

CEN EN 13606, *Health informatics -- Electronic Healthcare Record Communication*

3 Terms and definitions

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

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

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

[SOURCE: ISO 25720:2009(en), 4.1]

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

3.3
BSML
bioinformatic sequence markup language
extensible language specification and container for bioinformatic data

[SOURCE: ISO 25720:2009(en), 4.2]

3.4
Cell ML
cell markup language
a standard for representing and exchanging computer-based biological models

[SOURCE: ISO 25720:2009(en), 4.3]

3.5
CGAP
Cancer Gene Anatomy Project
genomic expression data collected for various tumorigenic tissues in both humans and mice.

Note to entry: CGAP also provides information on methods and reagents used in deriving the genomic data

[SOURCE: ISO 25720:2009(en), 4.4]

3.6
codon
a sequence of three nucleotides which together form a unit of genetic code in a DNA or RNA molecule.

3.7
dbSNP
database of SNPs (4.29) provided by the US National Center for Biotechnology Information (NCBI)

Note to entry: available at https://www.ncbi.nlm.nih.gov/SNP/

[SOURCE: ISO/TS 20428:2017(en), 3.9]

3.8
DICOM
digital imaging and communications in medicine
a standard in the field of medical informatics for exchanging digital information between medical imaging equipment (such as radiological imaging) and other systems, ensuring interoperability

[SOURCE: ISO 25720:2009(en), 4.6]

3.9
DNA
deoxyribonucleic acid
a molecule that encodes genetic information in the nucleus of cells

[SOURCE: ISO 25720:2009(en), 4.7]

3.10
DNA sequence variation
differences of DNA (4.8) sequence among individuals in a population

Note to entry: DNA sequence variation implies polymorphism (4.xx)

[SOURCE: ISO 25720:2009(en), 4.8]

3.11
DTD
document type definition
a document that contains formal definitions of all of the data elements in a particular type of HTML (4.15), SGML (4.28), or XML (4.38) document

[SOURCE: ISO 25720:2009(en), 4.9]

3.12
entry point
reference point that designate the class(es) from which the messages begin for the domain

[SOURCE: ISO 25720:2009(en), 4.10]

3.13
exon
any part of a gene that will encode a part of the final mature RNA produced by that gene after introns have been removed by RNA splicing.

3.14
gene-based medicine
medicine based on genes or genetic science

[SOURCE: ISO 25720:2009(en), 4.11]

3.15
GSVML
genomic sequence variation markup language
a standard for data exchange of genomic sequence variation data

[SOURCE: ISO 25720:2009(en)]

3.16
HTML
Hypertext Markup Language
a set of markup symbols or codes inserted in a file intended for display in a browser

[SOURCE: ISO 25720:2009(en), 4.12]

3.17
ICD-11
international classification of diseases 11th revision
a standard diagnostic tool for epidemiology, health management and clinical purposes

Note to entry: available at https://icd.who.int/

3.18
iCOS
clinical omics sub-information model for ICD

Note to entry: Add-on sub-information model to enhance the representation ability of ICD11 contents model to cover omics information.

3.19
intron
any nucleotide sequence within a gene that is removed by RNA splicing during maturation of the final RNA product

3.20
JPEG
joint photographic experts group
compression technique for images

[SOURCE: ISO 25720:2009(en), 4.13]

3.21
JSNP
Japanese single nucleotide polymorphisms
database of Japanese Single Nucleotide Polymorphisms

[SOURCE: ISO 25720:2009(en), 4.14]

3.22
markup language
ML
a set of symbols and rules for their uses when doing a markup of a document

[SOURCE: ISO 25720:2009(en), 4.15]

3.23
microarray gene expression markup language
MAGE-ML
a data format for describing information about DNA-array based experiments and gene expression data

3.24
neuro-ML
Neuro Markup Language
markup language (4.20) for describing models of neurons and networks of neurons.

[SOURCE: ISO 25720:2009(en), 4.16]

3.25
nroff
text-formatting program on Unix and unix-like systems

[SOURCE: https://en.wikipedia.org/wiki/Nroff]

3.26
omics
a field of study in biology ending in -omics

Note to entry: includes, but is not limited to, genomics, proteomics, and metabolomics.

3.27
pharmacogenomics
a branch of pharmaceutics aiming to develop rational means to optimize drug therapy, with respect to the patient's genotype

3.28
PolyMAPr
polymorphism mining and annotation programs
programs for polymorphism database mining, annotation, and functional analysis

[SOURCE: ISO 25720:2009(en), 4.19]

3.29
polymorphism
variation in the sequence of DNA (4.8) among individuals

Note to entry: polymorphism implies SNP (4.29) and STRP (4.32)

[SOURCE: ISO 25720:2009(en), 4.20]

3.30
RNA
ribonucleic acid
polymer of ribonucleotides occurring in a double-stranded or single-stranded form

[SOURCE: ISO 22174:2005, 3.1.3]

3.31
RNAML
a data format for exchanging RNA information

3.32
SBML
systems biology markup language
markup language (4.20) for simulations in systems biology

[SOURCE: ISO 25720:2009(en), 4.21]

3.33
SGML
standard generalized markup language
markup language (4.20) for document representation that formalizes markup and frees it of system and processing dependencies

[SOURCE: ISO 8879:1986, 4.305]

3.34
SNP
single nucleotide polymorphism
single nucleotide variation in a genetic sequence that occurs at appreciable frequency in the population

[SOURCE: ISO 25720:2009(en), 4.23]

3.35
SNOMED-CT
systematized nomenclature of medicine - Clinical Terms
dynamic, scientifically validated clinical health care terminology and infrastructure

[SOURCE: ISO 25720:2009(en), 4.24]

3.36
SOAP
simple object access protocol
lightweight protocol for exchange of information in a decentralized, distributed environment

[SOURCE: ISO 25720:2009(en), 4.25]

3.37
STRP
short tandem repeat polymorphism
variable segments of DNA (4.8) that are two to five bases long with numerous repeats

[SOURCE: ISO 25720:2009(en), 4.26]

3.38
troff
the major component of a document processing system developed by AT&T for the Unix operating system

3.39
VNTR
variable number of tandem repeat
class of polymorphism characterized by the highly variable copy number of identical or closely related sequences

[SOURCE: ISO 25720:2009(en), 4.28]

3.40
WML
wireless markup language
XML language used to specify content and user interface for WAP (Wireless Application Protocol) devices

[SOURCE: ISO 25720:2009(en), 4.29]

3.41
WGML
whole genome sequence markup language
markup language to represent complete genome sequence

3.42
XHTML
extensible HTML
hybrid between HTML (4.5) and XML (4.38) specifically designed for net device displays

[SOURCE: ISO 25720:2009(en), 4.30]

3.43
XML
Extensible Markup Language
pared-down version of SGML (4.28), designed especially for Web documents

[SOURCE: ISO 25720:2009(en), 4.31]

3.44
XML Schema
language for describing the structure and constraining the contents of XML documents

[SOURCE: ISO 25720:2009(en), 4.32]

4 OML specification

4.1 Specification requirements and OML positioning (informative)

The volume of omics data continues to increase. That data is stored in various databases in one of several different formats. In addition, annotative information relating real-world observations to the omics data adds to the data volume and data complexity. The volume of data, various structures and complexity have made the exchange of omics data difficult. While various markup languages have been used to represent omics data, but there has not been an omics-centric language.

OML is the first omics-centric markup language which is also oriented toward human health. To be useful for human healthcare, OML includes the capability to associate real-world observations with omics data. But the real-world observations can span a large spectrum of healthcare and may be expressed in other healthcare standards. As such, OML must be need harmonized with the other international standardization organizations such as Health Level Seven (HL7) and International Organization for Standardization (ISO).

In the current context, annotative information about omics is increasing and that information is embedding the information holes. The omics data itself is also increasing but is stored in various databases. The pitfall of omics data handling is the lack of standardization of the data formats for the organized omics. Historically, markup languages listed above have been used, and programs are developed to handle the omics information. However, there have been no omics centric markup languages so far. OML is the first omics centric markup language and is human health centric. Considering that omics has the great impact especially for human health and response, we can say that OML has the greatest potential to be the designated markup language for human healthcare. On the other hand, setting the applications to practical human health means it must handle direct or indirect annotations. Here the direct annotation indicates general annotative information such as omics associated other omics and experimental preparations. The indirect annotation indicates all of omics data and clinical data that result from omics data. To understand the omics based clinical situation of each patient, we need these kinds of additional information. Considering the requirements to add many kinds of additional information, the development and standardization of OML cannot stand alone and need harmonization with the other international standardization organizations such as Health Level Seven, an International Organization for Standardization.