INTERNATIONAL STANDARD

ISO 21393

First edition

Genomics informatics — Omics Markup Language (OML)

PROOF/ÉPREUVE

Reference number
ISO 21393:2021(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*, 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.[10] 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

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

[SOURCE: ISO 25720:2009, 4.2]

3.4
cancer genome anatomy project
CGAP
genomic expression data collected for various tumorigenic tissues in both humans and mice

Note 1 to entry: CGAP also provides information on methods and reagents used in deriving the genomic data

[SOURCE: ISO 25720:2009, 4.4, modified]

3.5
codon
sequence of three nucleotides which together form a unit of genetic code in a DNA or RNA molecule

3.6
dbSNP
database of *single nucleotide polymorphisms* (3.29) provided by the US National Center for Biotechnology Information (NCBI)

Note 1 to entry: Available at https://www.ncbi.nlm.nih.gov/SNP/.

[SOURCE: ISO/TS 20428:2017, 3.9]

3.7
digital imaging and communications in medicine
DICOM

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, 4.6]

3.8
DNA sequence variation
differences of DNA sequence among individuals in a population

Note 1 to entry: DNA sequence variation implies polymorphism 3.25.

[SOURCE: ISO 25720:2009, 4.8]

3.9
document type definition
DTD
document that contains formal definitions of all of the data elements in a particular type of *hypertext markup language* 3.13, *standard generalized markup language* (3.29), or *extensible markup language* (3.36) document

[SOURCE: ISO 25720:2009, 4.9]

3.10
entry point
reference point that designate the class(es) from which the messages begin for the domain

[SOURCE: ISO 25720:2009, 4.10, modified]

3.11
exon
part of a gene that will encode a part of the final mature RNA produced by that gene after *introns* (3.16) have been removed by RNA splicing

3.12
genomic sequence variation markup language
GSVML
standard for data exchange of genomic sequence variation data

3.13
hypertext markup language
HTML
set of markup symbols or codes inserted in a file intended for display in a browser

[SOURCE: ISO 25720:2009, 4.12, modified]

3.14
international classification of diseases
ICD
diagnose coding system for epidemiology, health management and clinical purposes

Note 1 to entry: ICD-10 is the 10th revision and ICD-11th is the 11th revision.

Note 2 to entry: available at https://icd.who.int/.

3.15
clinical omics sub-information model for international classification of diseases
clinical omics sub-information model for ICD
iCOS

sub-information model aiming to enhance the representation ability of ICD-11 contents model with covering omics information as an add-on part.

Note 1 to entry: Add-on sub-information model to enhance the representation ability of ICD-11 contents model to cover omics information.

3.16
intron
nucleotide sequence within a gene that is removed by RNA splicing during maturation of the final RNA product

3.17
joint photographic experts group
JPEG
compression technique for images

[SOURCE: ISO 25720:2009, 4.13]

3.18
markup language
ML
set of symbols and rules for their uses when doing a markup of a document

[SOURCE: ISO 25720:2009, 4.15]

3.19
microarray gene expression markup language
MAGE-ML
data format for describing information about DNA-array based experiments and gene expression data

3.20
neuro markup language
neuro-ML
markup language (3.18) for describing models of neurons and networks of neurons.

[SOURCE: ISO 25720:2009, 4.16]

3.21
nroff
unix text-formatting program that is a predecessor of the Unix *troff* (3.33) document processing system

[SOURCE: ISO 25720:2009, 4.17]

3.22
omics
field of study in biology ending in -omics

Note 1 to entry: It includes, but is not limited to, genomics, proteomics, and metabolomics.

3.23
pharmacogenomics
branch of pharmaceutics aiming to develop rational means to optimize drug therapy, with respect to the patient's genotype

3.24
polymorphism mining and annotation programs
PolyMAPr
programs for *polymorphism* (3.25) database mining, annotation, and functional analysis

[SOURCE: ISO 25720:2009, 4.19]

3.25
polymorphism
variation in the sequence of DNA among individuals

Note 1 to entry: Polymorphism implies single nucleotide polymorphism (3.29) and short tandem repeat polymorphism (3.32) .

[SOURCE: ISO 25720:2009, 4.20]

3.26
RNA markup language
RNAML
data format for exchanging RNA information

3.27
systems biology markup language
SBML
markup language (3.18) for simulations in systems biology

[SOURCE: ISO 25720:2009, 4.21]

3.28
standard generalized markup language
SGML
markup language (3.18) for document representation that formalizes markup and frees it of system and processing dependencies

[SOURCE: ISO 8879:1986, 4.305, modified]

3.29
single nucleotide polymorphism
SNP
single nucleotide variation in a genetic sequence that occurs at appreciable frequency in the population

[SOURCE: ISO 25720:2009, 4.23]

3.30
systematized nomenclature of medicine-clinical terms®[2)]
SNOMED-CT®
dynamic, scientifically validated clinical health care terminology and infrastructure

[SOURCE: ISO 25720:2009, 4.24]

3.31
simple object access protocol
SOAP
lightweight protocol for exchange of information in a decentralized, distributed environment

[SOURCE: ISO 25720:2009, 4.25]

3.32
short tandem repeat polymorphism
STRP
variable segments of DNA that are two to five bases long with numerous repeats

[SOURCE: ISO 25720:2009, 4.26]

3.33
troff
major component of a document processing system developed by AT&T for the Unix operating system

3.34
wireless markup language
WML
extensible markup language used to specify content and user interface for WAP (Wireless Application Protocol) devices

[SOURCE: ISO 25720:2009, 4.29]

3.35
extensible HTML
XHTML
hybrid between *hypertext markup language* 3.13 and *extensible markup language* (3.36) specifically designed for net device displays

[SOURCE: ISO 25720:2009, 4.30]

3.36
extensible markup language
XML
pared-down version of *standard generalized markup language* (3.29), designed especially for web documents

[SOURCE: ISO 25720:2009, 4.31]

2) SNOMED CT is the registered trademark of International Health Terminology Standards Development Organisation. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of the product named.

3.37
XML schema
language for describing the structure and constraining the contents of extensible markup language documents

[SOURCE: ISO 25720:2009, 4.32]

4 OML specification

4.1 Specification requirements and OML positioning

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 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, it can be said 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 shall handle direct or indirect annotations. Here the direct annotation shall indicate general annotative information such as omics associated other omics information and experimental preparations. The indirect annotation shall indicate all of omics data and clinical data that result from omics data. To understand the omics based clinical situation of each patient, these kinds of additional information is required. Considering the requirements to add many kinds of additional information, the development and standardization of OML cannot stand alone and shall need harmonization with the other documents from the other international standardization organizations.

OML intends to be used in data exchange messages related to human health. In development and standardization of OML in this application domain, keeping an eye on the patient safety, the clinical efficiency, and the medical costs shall always be required. For the patient safety from an informational side, the conservation and the protection of patient information shall be deemed important. For the enhancement of the clinical efficiency, the simplicity and the easy understandability shall be deemed important. For the medical cost reduction, the adaptation ability and installation ease shall be deemed important.

OML tries to respond to these basic requirements by providing the sharable XML based data exchanging format. OML can be used for the clinically omics data exchange among various types of data formats. In the greater framework of clinical data standardization, OML shall play a part of describing the omics data and its necessary information.

4.2 OML Structure

A valid OML expression shall be structured in accordance with the following, also see Figure 2:

— The outline structure of OML is shown in Figure 2.

OML shall consist of three data criteria:

— omics data;

— direct annotation;

— indirect annotation.

The omics data criterion shall describe, for each omics

the straight forward omics data as:

— type;

— position;

— length;

— egion;

— etc.

The direct annotation criterion shall describe, for each omics

the attached data of omics data as:

— experiment analysis;

— epidemiology;

— associated omics;

— etc.

The indirect annotation criterion shall describe

the explanatory/higher-level information of omics data as:

— the clinical information;

— the environmental data.

These data criteria shall have relations to each other internally.

— The detailed structure of OML shall be given as in Figures 3 to 23.

4.3 OML DTD and XML Schema

The DTD of OML is available for information at https://standards.iso.org/iso/21393/ed-1/en.

The XML schema of OML used is available at https://standards.iso.org/iso/21393/ed-1/en.

5 OML development process

Step 1: Set the elements and needs according to the investigated use cases including use case with WHO ICD-11 iCOS use.

Step 2: Construct the basic structure and DTD

Step 3: Investigate the existing biological ML, in particular GSVML (ISO 25720), and its applicability to the needs. (Comparison with MAGE-ML, BSML, SBML, RNAML,[21] ProML, CellML, PolyMAPr)

Step 4: Refine the basic structure and DTD, Construct the XML Schema (XSD)

Step 5: Investigate the existing format (their data format comparison).

Step 6: Check the interface ability to the Health Level Seven® Models.

Step 7: Redefine the needs to OML and its demanded elements.

Step 8: Refine the basic structure, DTD, and XML Schema

Figure 24 outlines of the process of the development. The design work was done in harmony with HL7® Clinical Genomics WG, CDISC BRIDGE group, WHO FIC ITC group for both ISO 25720 (GSVML) and this document. There were "to and fro" processes between design work and the standardization process.

Additionally, the interface between OML, ISO 13606 (all parts), and SNOMED-CT® is analyzed.

Additional informative input to the development of this document are included in Annex A.

6 Figures

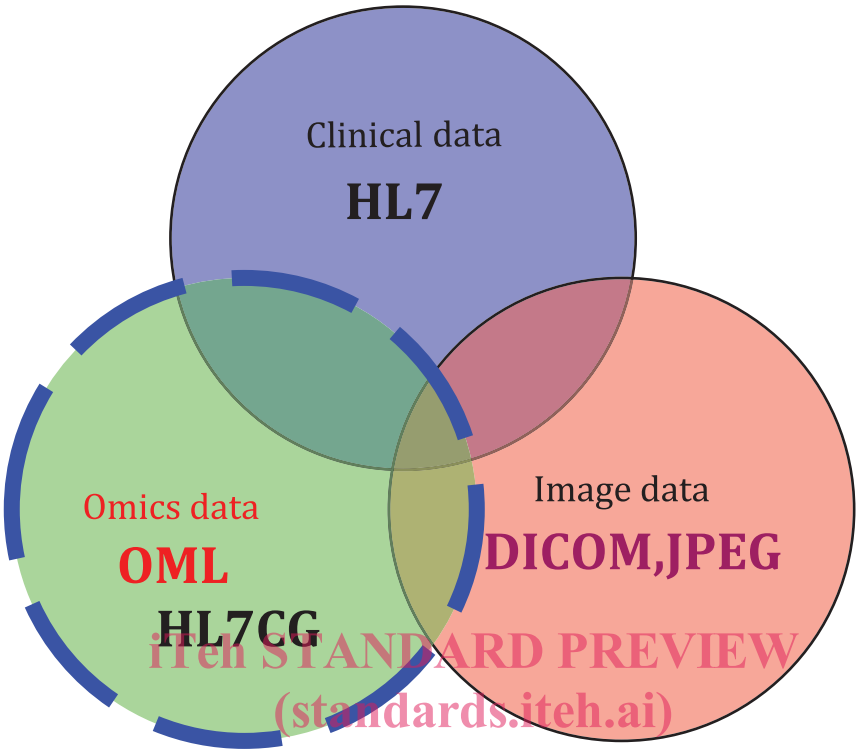

Figure 1 — Major data types of health care

In the current electronic network world, there are multiple different types of data for healthcare as shown in Figure 1. Besides clinical data and image data, as moving into this next generation post genomic era, overwhelming amounts of omics data is creating internationally. Standards organizations are developing standards for these data; Health Level Seven® (HL7®) develops standards for clinical data, DICOM and JPEG develop standards for image data; and Omics Markup Language (OML) defines a standard for omics data, especially human-related omics data. The core target for the OML is the data exchange format.

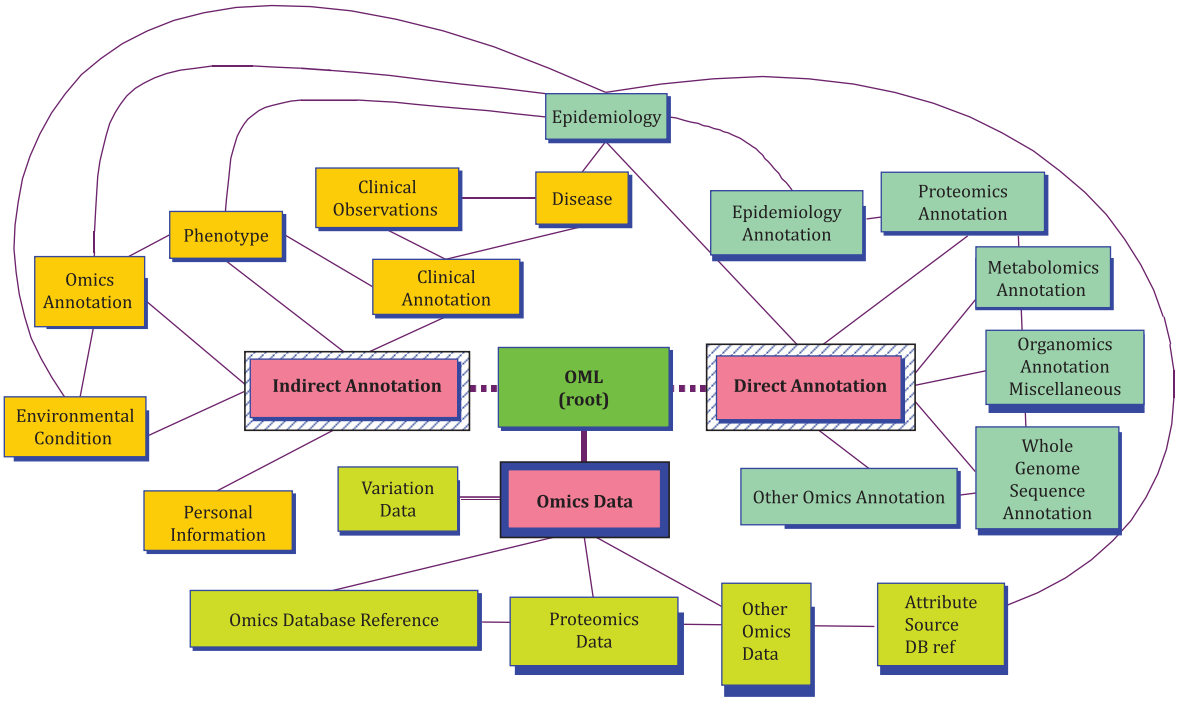

Figure 2 — The outlined structure of OML

The overall structure of OML is centred on the OML (root) and the Omics data content – either Variation, Proteomics, or other Omics data. Information related to omics processes, or otherwise not included in Omics data are contained in Direct Annotations. Indirect Annotations permit related clinical, phenotypic, environmental, and similar information to be included in the OML document.

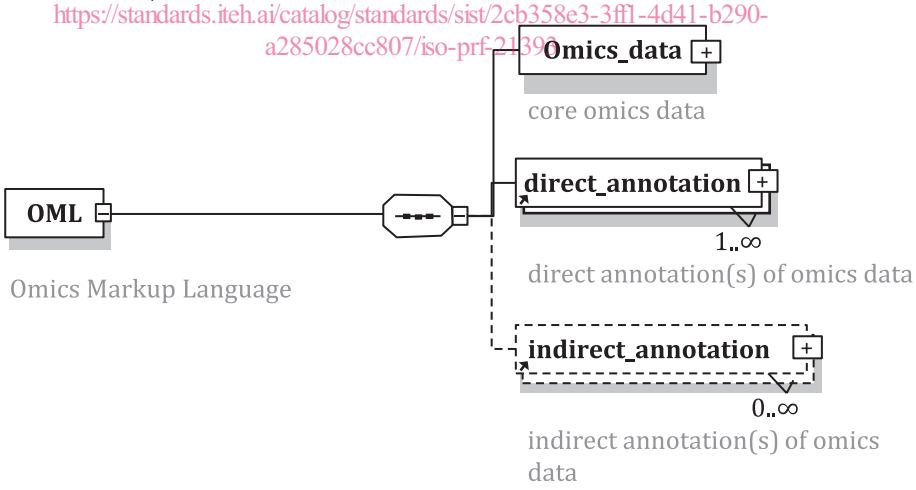

Figure 3 — Detailed structure of OML: OML root (OML)

Figure 3 shows to overall structure of an OML document instance, where the OML root element (OML) is composed of omics_data, direct_annotation, and indirect_annotation.

Further description of omics_data can be found in and following Figure 4.

Further description of direct_annotation can be found in and following Figure 11.

Further description of indirect_annotation can be found in and following Figure 18.