INTERNATIONAL STANDARD

ISO 25720

First edition
2009-08-15

Health informatics — Genomic Sequence Variation Markup Language (GSVML)

Informatique de santé — Langage de balisage de la variation de séquence génomique

Reference number
ISO 25720:2009(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword iv
Introduction v
1 Scope 1
2 Conformance 1
2.1 Purpose 1
2.2 Needs and general requirements 1
3 Normative references 2
4 Terms and definitions 2
5 GSVML specification 5
5.1 Specification requirements and GSVML positioning 5
5.2 GSVML structure 5
5.3 GSVML DTD and XML schema 5
6 GSVML development process 5
Annex A (normative) **DTD of GSVML** 22
Annex B (normative) **XML schema of GSVML** 46
Annex C (informative) **Basic reference works** 105
Bibliography 131

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.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

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.

ISO 25720 was prepared by Technical Committee ISO/TC 215, *Health informatics*.

Introduction

In the current electronic world, there are multiple different types of data for healthcare, as shown in Figure 1. Besides clinical data and image data, as we move into this post genomic era, we are creating, internationally, overwhelming amounts of genomic data. The International Standards developing organizations are developing standards for these data; Health Level Seven develops standards for clinical data, DICOM and JPEG develop standards for image data. Genomic Sequence Variation Markup Language (GSVML) defines a standard for genomic data, especially human-related DNA variation data. The core target for the GSVML is the Single Nucleotide Polymorphism (SNP).

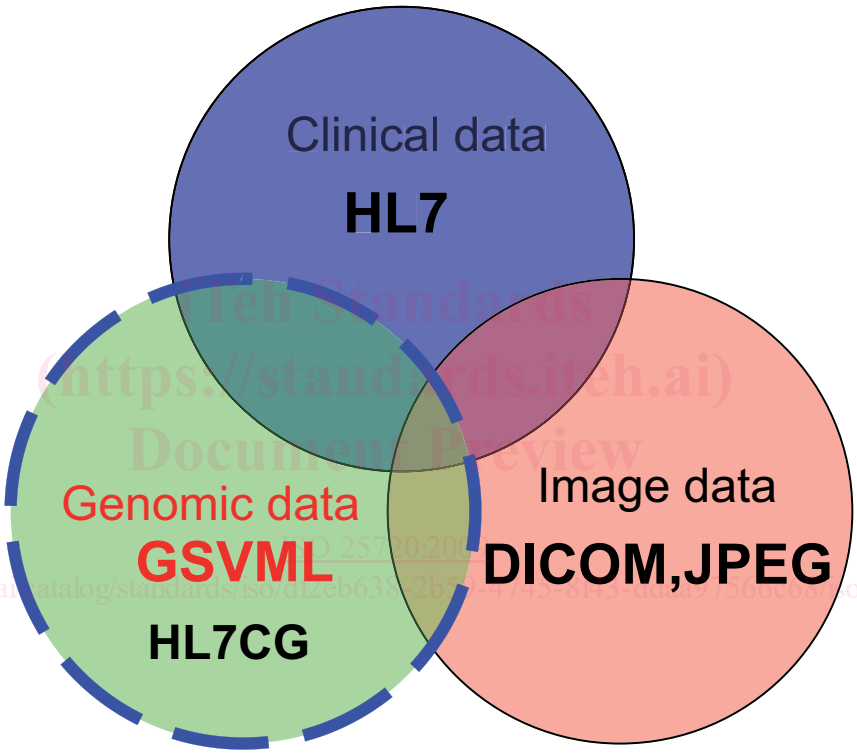


Figure 1 — Major data types of health care

In this post genomic era, the management of health-related data is becoming increasingly important to both genomic research and genome-based medicine (see reference [1]). Informational approaches to the management of clinical, image and genomic 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 -omics data. Genomic data include genome sequence, DNA sequence variation and other genome-based data such as expression data, proteomics data, molecular network, etc. As an entry point, this International Standard focuses on the DNA sequence variation. Among the DNA sequence variation, the SNP is selected as the core object because of the following three reasons.

a) SNP is the most researched sequence variation for human health.

b) In the current context, vast amounts of SNP data exist around the world in various types of data formats. As a result of the recent explosion in SNP research, the vast amounts of experimental data have been

accumulating in many databases in various types of data formats. These data await utilization in drug discovery, clinical diagnosis and clinical research.

c) SNP data already have a great impact for human applications such as gene-based medicine and pharmacogenomics.

With a view to this context, the international community requires an interoperable format for the interchange of SNP data. Prior to the standardization development, we elucidated the need for data exchange among the human health-related facilities that have various types of data formats.

In the present circumstances, SNP is expected to be a key to understanding human response to external stimuli such as any kind of alien invasions, therapies, and the environmental interactions (see reference [2]). Bacterial infection is an example of alien invasion, and the responses to the infections are different amongst individuals. According to the therapy, the side effects to a drug are different amongst the patients. These responses are also different in various environments.

The Markup Language is a set of symbols and rules for their use when doing a markup of a document (see reference [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 (see reference [6]). XML acts as the basis for Extensible HTML (XHTML) [7] and Wireless Markup Language (WML) (see reference [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 (see reference [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) (see reference [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 databases such as dbSNP [16], the Cancer Gene Anatomy Project (CGAP) (see reference [17]), and Japanese single nucleotide polymorphisms (JSNP) (see reference [18]) through programming.

To utilize the accumulated SNP data among many facilities around the world, standards for the interchange of SNP 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 genomic data message handling, Health Level Seven Clinical Genomics Special Interest Group [19] has summarized clinical use cases for general genomic data. The GSVML 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 International Standard elucidates the needs and the requirements for GSVML and then proposes the specification of GSVML for the international standardization.

Health informatics — Genomic Sequence Variation Markup Language (GSVML)

IMPORTANT — The electronic file of this document contains colours which are considered to be useful for the correct understanding of the document. Users should therefore consider printing this document using a colour printer.

1 Scope

This International Standard is applicable to the data exchange format that is designed to facilitate the exchange of the genomic sequence variation data around the world, without forcing change of any database schema. From an informatics perspective, GSVML defines the data exchange format based on XML. The scope of this International Standard is the data exchange format, but the database schema itself is outside the scope of this International Standard. From a biological point of view, all genetic sequence variations are taken into consideration and are within the scope of this International Standard, while polymorphisms, especially SNPs, are the main focus of this International Standard. In other words, the annotations of variation as clinical concerns and -omics concerns are within the scope of this International Standard. Though SNPs exist in various biological species, the scope of this International Standard covers the human health associated species as human, cell line, and preclinical animals. The other biological species are outside the scope of this International Standard. The clinical field is within the scope of this International Standard, but the basic research fields and other scientific fields are outside the scope of this International Standard. Here, clinical research, including drug discovery, is within the scope of this International Standard. As for supposed application fields, the main focus is in human health, including clinical practice, preventive medicine, translational research and clinical researches.

2 Conformance

2.1 Purpose

This International Standard provides a data exchange format for genomic sequence variation data in human health. This International Standard provides the GSVML specification mainly for the case of SNP and Short Tandem Repeat Polymorphism (STRP). Considering that SNP and STRP are the major and simple polymorphisms in human health research, centering on them and expanding the specification to the other sequence variation data seems reasonable. This International Standard allows for the expandability of GSVML from SNP and STRP to other sequence variation data.

2.2 Needs and general requirements

The vast volume of experimental data from the recent explosion of genomic sequence variation research has produced an overwhelming amount of data stored in many databases with various types of format worldwide. Standardization of data exchange is urgent for managing, analysing, and utilizing these data. Standardizing the interoperable format is necessary for easy and convenient genomic sequence variation data exchange. Considering that genomic sequence variation, especially SNP and STRP, has its significant meaning in the gene-based medicine and the pharmacogenomics for human health, the data exchange format is the key to enhancing the gene-based clinical research and the gene-based medicine.

The management of genomic data is as critical as the 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 these genomic data for human health. In order to use these data effectively and efficiently, standards must be developed to permit the interoperable interchange of genomic 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 International Standard addresses those requirements, using a Markup Language.

3 Normative references

The following referenced documents are indispensable for the application 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.

EN 13606 (all parts), *Health informatics — Electronic healthcare record communication*

4 Terms and definitions

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

4.1
actor
something or someone who supplies a stimulus to the system

NOTE Actors include both humans and other quasi-autonomous things, such as machines, computer tasks and systems.

4.2
Bioinformatic Sequence Markup Language
BSML
extensible language specification and container for bioinformatic data

4.3
Cell Markup Language
Cell ML
Extensible Markup Language to provide a standard method for representing and exchanging computer-based biological models

4.4
Cancer Gene Anatomy Project
CGAP
database containing genomic expression data collected for various tumorgenic tissues in both humans and mice and also providing information on methods and reagents used in deriving the genomic data

4.5
dbSNP
database of SNPs provided by the US National Center for Biotechnology Information (NCBI)

4.6
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

4.7
deoxyribonucleic acid
DNA
molecule that encodes genetic information in the nucleus of cells

4.8
DNA sequence variation
differences of DNA sequence among individuals in a population

NOTE DNA sequence variation implies **polymorphism** (4.20).

4.9
Document Type Definition
DTD
separate document that contains formal definitions of all of the data elements in a particular type of HTML, SGML or XML document

4.10
entry point
reference point that designate the class(es) from which the messages begin for the particular domain

4.11
gene-based medicine
medicine based on genes or genetic science

4.12
Hypertext Markup Language
HTML
set of markup symbols or codes inserted in a file intended for display on a World Wide Web browser page

4.13
Joint Photographic Experts Group
JPEG
compression technique for images

4.14
Japanese single nucleotide polymorphisms
JSNP
database of Japanese single nucleotide polymorphisms

4.15
markup language
ML
set of symbols and rules for their uses when doing a markup of a document

4.16
Neuro Markup Language
Neuro-ML
markup language for describing models of neurons and networks of neurons

4.17
nroff
unix text-formatting program that is a predecessor of the Unix troff document processing system

4.18
pharmacogenomics
branch of pharmaceutics aiming to develop rational means to optimize drug therapy, with respect to the patient's genotype

4.19
Polymorphism Mining and Annotation Programs
PolyMAPr
programs for polymorphism database mining, annotation and functional analysis

4.20
polymorphism
variation in the sequence of DNA among individuals

NOTE Polymorphism implies **SNP** (4.23) and **STRP** (4.26).

4.21
Systems Biology Markup Language
SBML
markup language for simulations in systems biology

4.22
Standard Generalized Markup Language
SGML
standard for defining description of the structure of different types of electronic documents

4.23
Single Nucleotide Polymorphism
SNP
single nucleotide variation in a genetic sequence that occurs at appreciable frequency in the population

4.24
Systematized Nomenclature of Medicine – Clinical Terms
SNOMED CT
dynamic, scientifically validated clinical health care terminology and infrastructure

4.25
Simple Object Access Protocol
SOAP
lightweight protocol for exchange of information in a decentralized, distributed environment

4.26
Short Tandem Repeat Polymorphism
STRP
variable segments of DNA that are two to five bases long with numerous repeats

4.27
troff
document processing system developed by AT&T for the Unix operating system

4.28
variable number of tandem repeat
VNTR
class of polymorphism characterized by the highly variable copy number of identical or closely related sequences

4.29
Wireless Markup Language
WML
XML language used to specify content and user interface for WAP (wireless application protocol) devices

4.30
Extensible HTML
XHTML
hybrid between HTML and XML specifically designed for net device displays

4.31
Extensible Markup Language
XML
pared-down version of SGML, designed especially for web documents

4.32
XML schema
language for describing the structure and constraining the contents of XML documents

5 GSVML specification

5.1 Specification requirements and GSVML positioning

In the current context, annotative information about genomic sequence variation is increasing, which is filling in the gaps in information. The genomic sequence variation data themselves are also increasing but are stored in various databases. This trend is typical of SNP data. The pitfall of genomic sequence variation data handling is the lack of standardization of the data formats for the genomic sequence variation. Historically, the markup languages listed in Clause 4 have been used, and programs are developed to handle the genomic information. However, there have been no genomic sequence variation centric markup languages so far. GSVML is the first genomic sequence variation centric markup language and is human health centric. Considering that SNP is a highly researched polymorphism and has a great impact, especially for human health and response, we can say that GSVML 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 SNP annotations. Here the direct SNP annotation indicates general annotative information such as SNP associated genes and experimental preparations. The indirect SNP annotation indicates all of the -omics data and clinical data that result from SNP variation. To understand the gene-based clinical situation of each patient, we need this kind of additional information. Considering the requirement to add many kinds of additional information, the development and standardization of GSVML cannot stand alone and need harmonization with the other international standardization organizations such as Health Level Seven.

GSVML is intended to be used in data exchange messages related to human health. In the development and standardization of GSVML in this application domain, we must always keep an eye on the patient's safety, clinical efficiency and medical costs. For the patient's safety, from an informational viewpoint, the conservation and the protection of patient information are important. For the enhancement of clinical efficiency, simplicity and ease of understanding are important. For medical cost reduction, the adaptation ability and ease of installation are important. GSVML tries to respond to these basic requirements by providing the sharable XML based data exchanging format. GSVML can be used for the clinically genomic sequence variation data exchange among various types of data formats. In the greater framework of clinical data standardization, GSVML plays the part of describing the genomic sequence variation data and their necessary information.

5.2 GSVML structure

The outlined structure of GSVML is shown in Figure 3. GSVML consists of three data criteria, *viz.* variation data, direct annotation and indirect annotation. The variation data criterion describes the straightforward variation data as allele, type, position, length, region, etc. The direct annotation criterion describes the attached data of variation data as experiment analysis, epidemiology or associated gene, etc. The indirect annotation criterion describes the explanatory/higher-level information of variation data such as the -omics data, the clinical information and the environmental data. These data criteria have internal relations to each other. The detailed structure of GSVML is shown in Figures 4 to 21.

5.3 GSVML DTD and XML schema

The DTD of GSVML is shown in Annex A. The XML schema of GSVML is shown in Annex B.

6 GSVML development process

The development of GSVML followed eight steps:

— Step 1: Set the elements and needs according to the investigated use cases.

We prepared six use cases for three typical criteria. Four use cases concerned the clinical practice, and one use case for each clinical trial and translational research.

— Step 2: Construct the basic structure and DTD.

— Step 3: Investigate the existing biological ML, 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 SNP databases (their data format comparison).

— Step 6: Check the interface ability to the Health Level Seven Genotype Model.

— Step 7: Redefine the needs of GSVML and its demanded elements.

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

Figure 2 shows the outline of the process of the development. We did design work in harmony with HL7 Clinical Genomics SIG. There were “to and fro” processes between design work and the standardization process.

Additionally, we analysed the interface between GSVML and EN 13606, SNOMED-CT.

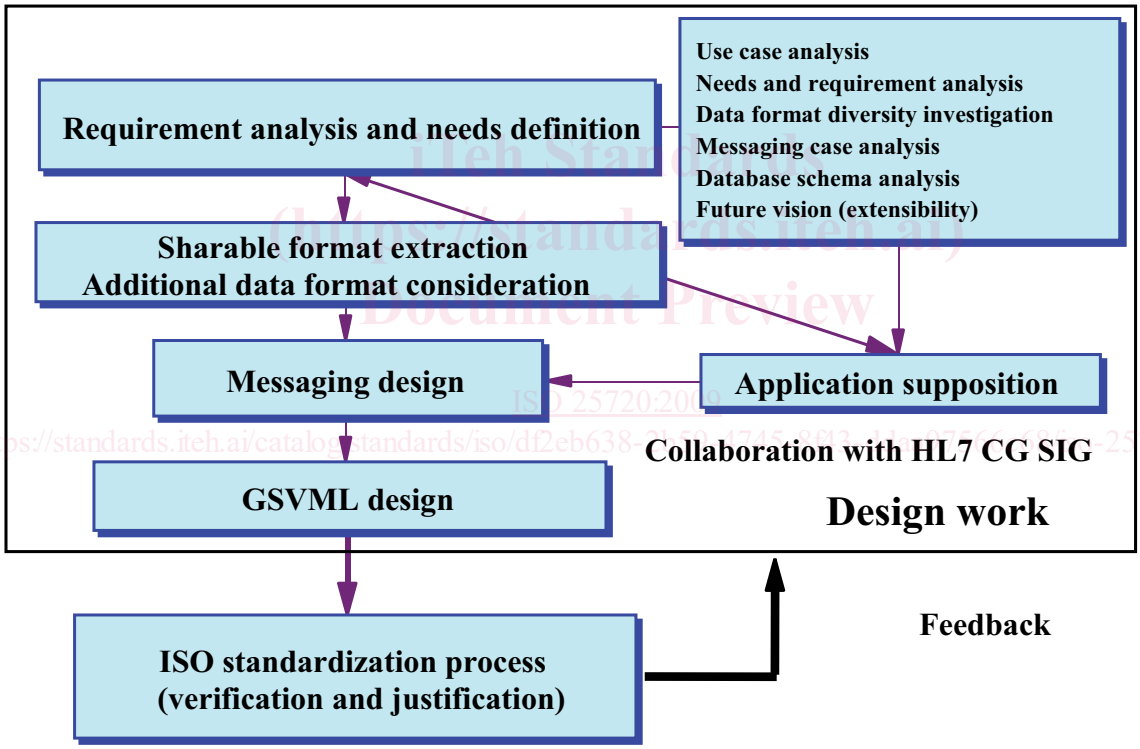
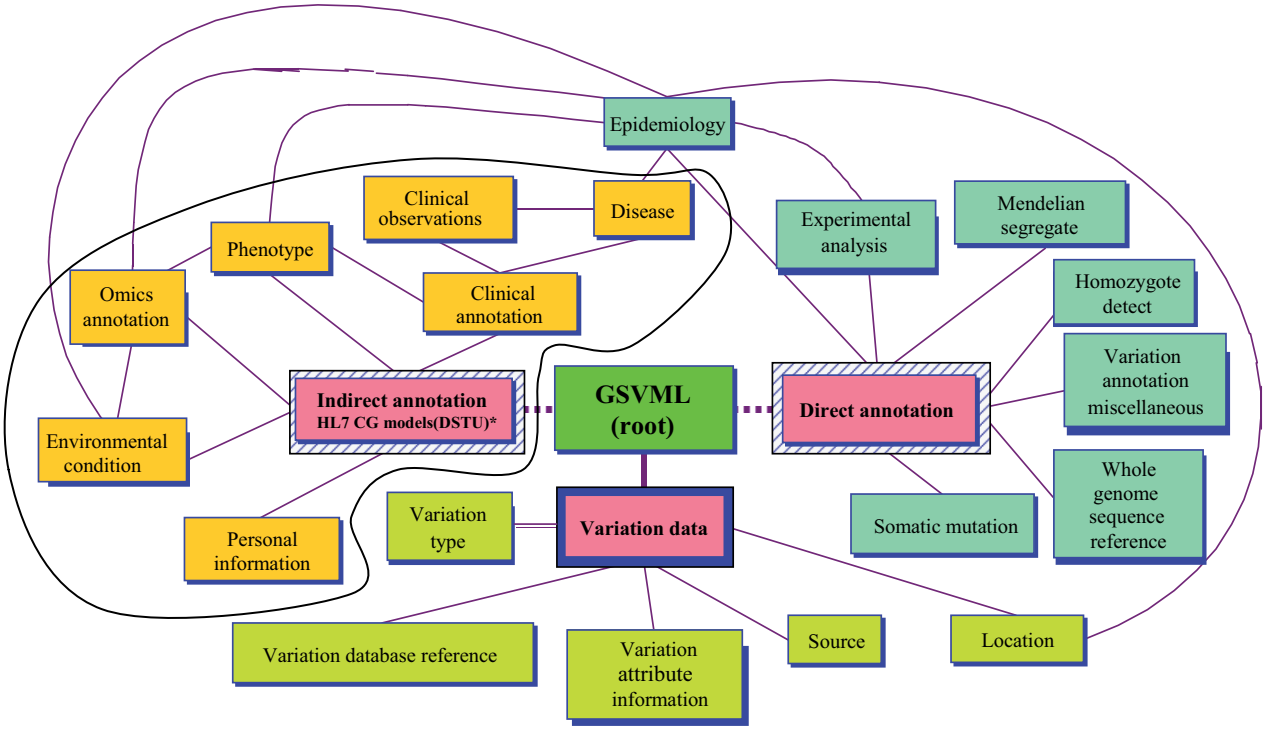


Figure 2 — Outline of the process of GSVML development



* HL7 CG models (DSTU) will be used instead of indirect annotation criterion.

Figure 3 — The outlined structure of GSVML

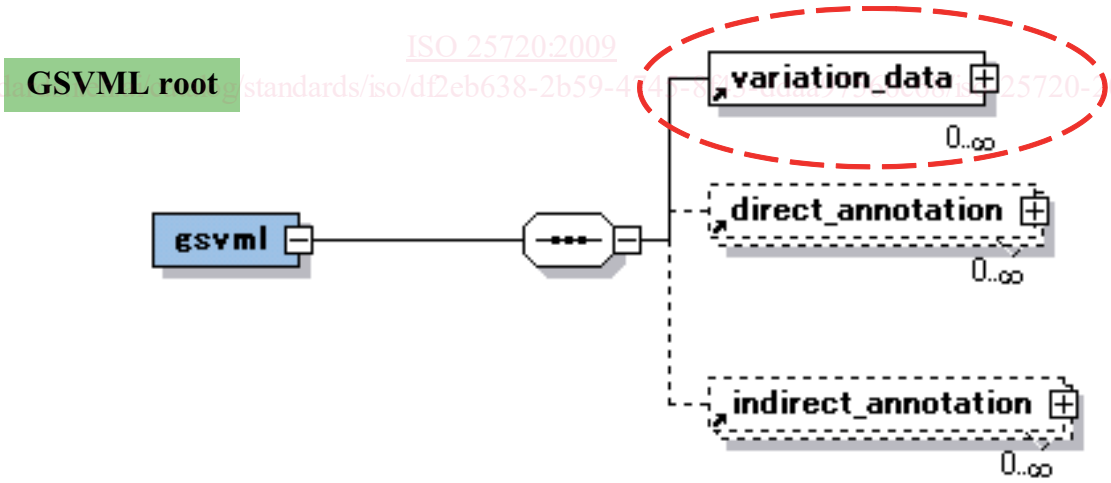


Figure 4 — Detailed structure of GSVML

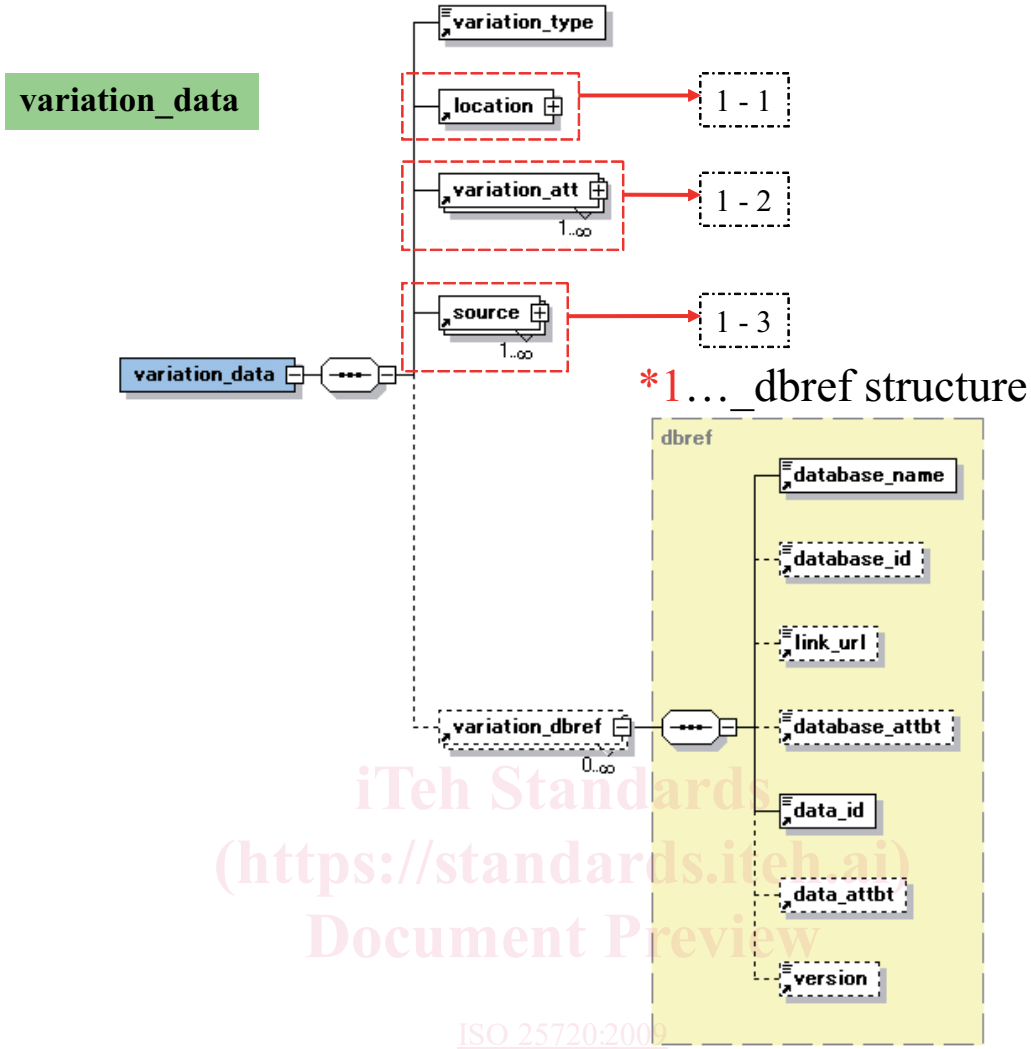


Figure 5 — Detailed structure of GSVML