PROJET DE NORME INTERNATIONALE

ISO/DIS 21393

ISO/TC **215**	Secrétariat: **ANSI**
Début de vote: **2019-07-16**	Vote clos le: **2019-10-08**

Informatique de santé — Langage de balisage Omics (OML)

Health informatics — Omics Markup Language (OML)

ICS: 35.240.80

Le présent document est distribué tel qu'il est parvenu du secrétariat du comité.

TRAITEMENT PARALLÈLE ISO/CEN

Numéro de référence
ISO/DIS 21393:2019(F)

ISO copyright office
Case postale 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Tél.: +41 22 749 01 11
Fax: +41 22 749 09 47
E-mail: copyright@iso.org
Website: www.iso.org

Publié en Suisse

Sommaire

Page

Avant-propos

L'ISO (Organisation internationale de normalisation) est une fédération mondiale d'organismes nationaux de normalisation (comités membres de l'ISO). L'élaboration des Normes internationales est en général confiée aux comités techniques de l'ISO. Chaque comité membre intéressé par une étude a le droit de faire partie du comité technique créé à cet effet. Les organisations internationales, gouvernementales et non gouvernementales, en liaison avec l'ISO participent également aux travaux. L'ISO collabore étroitement avec la Commission électrotechnique internationale (IEC) en ce qui concerne la normalisation électrotechnique.

Les procédures utilisées pour élaborer le présent document et celles destinées à sa mise à jour sont décrites dans les Directives ISO/IEC, Partie 1. Il convient, en particulier de prendre note des différents critères d'approbation requis pour les différents types de documents ISO. Le présent document a été rédigé conformément aux règles de rédaction données dans les Directives ISO/IEC, Partie 2 (voir www.iso.org/directives).

L'attention est attirée sur le fait que certains des éléments du présent document peuvent faire l'objet de droits de propriété intellectuelle ou de droits analogues. L'ISO ne saurait être tenue pour responsable de ne pas avoir identifié de tels droits de propriété et averti de leur existence. Les détails concernant les références aux droits de propriété intellectuelle ou autres droits analogues identifiés lors de l'élaboration du document sont indiqués dans l'Introduction et/ou dans la liste des déclarations de brevets reçues par l'ISO (voir www.iso.org/brevets).

Les appellations commerciales éventuellement mentionnées dans le présent document sont données pour information, par souci de commodité, à l'intention des utilisateurs et ne sauraient constituer un engagement.

Pour une explication de la nature volontaire des normes, la signification des termes et expressions spécifiques de l'ISO liés à l'évaluation de la conformité, ou pour toute information au sujet de l'adhésion de l'ISO aux principes de l'Organisation mondiale du commerce (OMC) concernant les obstacles techniques au commerce (OTC), voir le lien suivant : www.iso.org/iso/fr/foreword.html.

Le présent document a été élaboré par le comité technique ISO/TC 215, *Informatique de santé*, sous-comité SC 1, *Génomique clinique*.

Il convient que l'utilisateur adresse tout retour d'information ou toute question concernant le présent document à l'organisme national de normalisation de son pays. Une liste exhaustive desdits organismes se trouve à l'adresse www.iso.org/members.html.

Introduction

Dans cette ère post-génomique de nouvelle génération, la gestion des données de santé devient de plus en plus importante tant pour la médecine omique (« omics ») que pour la médecine basée sur les approches omiques [1]. Les approches informationnelles de la gestion des données cliniques, d'images et d'omique commencent à avoir autant de valeur que les recherches ordinaires en laboratoire. Il existe aujourd'hui de nombreux types de données omiques de par le monde qui attendent une utilisation efficace dans le domaine de la santé humaine. Pour atteindre cet objectif, le premier obstacle à franchir est de développer un format de données et des normes de message pour prendre en charge l'échange de données omiques cliniques. Les données omiques comprennent la séquence omique, la variation de séquence et d'autres données d'expression, les données protéomiques, le réseau moléculaire, etc. Comme point d'entrée, la présente norme se concentre sur l'échange de données.

Dans les circonstances actuelles, on s'attend à ce que l'omique soit une clé pour comprendre la réponse humaine aux stimuli externes tels que n'importe quels types d'invasions étrangères, de thérapies, et d'interactions environnementales [2]. L'infection bactérienne est un exemple d'invasion étrangère et les réponses aux infections diffèrent d'un individu à l'autre. Selon la thérapie utilisée, les effets secondaires d'un médicament diffèrent d'un patient à l'autre. Ces réponses diffèrent également d'un environnement à l'autre. Le nombre de ces recherches omiques ayant explosé récemment, les données expérimentales s'accumulent en grande quantité dans de nombreuses bases de données sous différents types de formats de données. Ces données attendent d'être utilisées dans la découverte de médicaments, le diagnostic clinique et les recherches cliniques.

Le langage de balisage est un ensemble de symboles et de règles permettant de les utiliser dans le balisage d'un document [3]. Le premier langage de balisage normalisé a été le SGML (langage normalisé de balisage généralisé) [4] qui a de fortes similitudes avec les langages de présentation de texte troff et nroff qui accompagnent les systèmes Unix. Le langage HTML (langage de balisage hypertexte) est basé sur SGML [5]. XML (langage de balisage extensible) est une version réduite du SGML, conçue particulièrement pour les documents Web [6]. XML sert de base au XHTML (HTML extensible) [7] et au WML (langage de balisage sans fil) [8] ainsi qu'à des définitions normalisées d'interaction système telles que SOAP (Simple Object Access Protocol) [9]. Par contre, la présentation de texte ou la sémantique est souvent définie sous une forme purement interprétable par machine, comme dans la plupart des formats de fichiers de traitement de texte [10].

Ces dernières années, le langage de balisage dans le domaine biomédical basé sur XML connaît un développement soutenu afin de renforcer l'échange de données parmi des chercheurs. Le BSML (langage de balisage de séquence bioinformatique) [11], le SBML (langage de balisage en biologie des systèmes) [12], le Cell ML (langage de balisage de cellules) [13] et le Neuro-ML (Langage de balisage neuronal) [14] sont des exemples de langages de balisage. Le Polymorphism Mining and Annotation Programs (PolyMAPr) [15] est centré sur le SNP et tente de réaliser l'exploration, l'annotation et l'analyse fonctionnelle des bases de données publiques telles que dbSNP [16], CGAP [17], et JSNP [18] par le biais de la programmation. Le langage de balisage de variation de la séquence génomique (GSVML) de l'ISO 25720 est le premier langage de balisage normalisé pour l'échange de données relatives à la variation de la séquence génomique dans un contexte clinique.

Le langage de balisage Omics (OML) vise à fournir le format normalisé d'échange de données pour les sciences omiques dans le domaine de la santé humaine.

L'essor récent de la recherche omique a généré d'importantes quantités de données conservées dans de nombreuses bases de données sous différents formats. La gestion, l'analyse et l'utilisation de ces données exigent une normalisation de l'échange de données. Compte tenu de l'importance des sciences omiques pour la médecine moléculaire et la pharmacogénomique, en particulier la transcriptomique, la protéomique, la signalomique et la métabolomique, le format d'échange de données est essentiel pour améliorer la recherche clinique et la médecine basées sur des approches omiques.

Les approches informationnelles ont récemment gagné en importance tant pour la recherche omique que pour la médecine basée sur les sciences omiques. Dans cette nouvelle ère, la gestion des données omiques est devenue aussi essentielle que celle des données de recherche fondamentale. Il existe de nombreux types de données omiques dans le monde et le temps est venu d'utiliser efficacement ces données pour la santé humaine. Pour utiliser ces données de manière efficace et efficiente, il est impératif d'élaborer des normes pour permettre l'échange interopérable des données omiques dans le monde. Ces normes doivent définir le format de données ainsi que les messages à utiliser pour échanger et partager ces données à l'internationale. La présente norme répond à ces exigences à l'aide d'un langage de balisage.

OML est un cadre de base pour tous les types de données omiques cliniques. Chaque catégorie du domaine omique sera présentée sous la forme d'une composante complémentaire spécifique. Par exemple, le langage de balisage du séquençage de génome complet formera une composante complémentaire spécifique pour des données de séquençage de génome complet, et le langage de balisage de la variation de la séquence génomique formera une composante complémentaire spécifique des données de variation de la séquence génomique.

Pour utiliser les données omiques cumulées parmi de nombreux établissements à travers le monde, des normes doivent être définies autour de l'échange de données omiques. Les normes requises incluent la définition d'un format de données et de messages d'échange. Le langage de balisage est le choix raisonnable pour répondre à ce besoin. Quant à la gestion des messages de données omiques, le groupe de travail de génomique clinique au sein du Health Level Seven [19] a récapitulé les cas d'utilisation clinique pour les données omiques générales. Le projet OML a contribué à ces efforts. En outre, ces travaux ont incorporé des cas d'utilisation basés sur le « Millennium Project » japonais [20]. Basé sur ces contextes et investigations, le présent document élucide les besoins et les exigences pour l'OML et propose ensuite la spécification de l'OML en vue de la normalisation internationale.

Une liste de références se rapportant à la présente partie de l'ISO/DIS 21393 est donnée dans la bibliographie.

Informatique de santé — Langage de balisage Omics (OML)

1 Domaine d'application

OML est un format d'échange de données conçu pour faciliter l'échange de données omiques à travers le monde sans introduire de modifications aux bases de données existantes.

D'un point de vue informatique, OML est un format d'échange de données basé sur XML. Le format d'échange de données (par exemple, schéma et DTD XML) entre dans le domaine d'application. La structure des systèmes et des bases de données qui envoient ou reçoivent les schémas d'informations ne s'inscrivent pas dans le domaine d'application.

D'un point de vue biologique, tous les types d'omique relèvent du domaine d'application, mais les détails (par exemple, détails des variations de la séquence génomique ou séquence génomique complète) en sont exclus. Les annotations incluant les questions cliniques et les relations avec les autres questions omiques entrent dans le domaine d'application.

L'application est concentrée sur la santé humaine, y compris les pratiques cliniques, la médecine préventive, la recherche translationnelle et la recherche clinique, notamment la découverte de médicaments. Le domaine d'application couvre les espèces associées à la santé humaine, notamment l'homme, les animaux en préclinique et les lignées cellulaires associées. Les autres espèces, recherches fondamentales et autres domaines scientifiques ne relèvent pas du domaine d'application.

2 Références normatives

Les documents suivants cités dans le texte constituent, pour tout ou partie de leur contenu, des exigences du présent document. Pour les références datées, seule l'édition citée s'applique. Pour les références non datées, la dernière édition du document de référence s'applique (y compris les éventuels amendements).

ISO 25720:2009, *Informatique de santé — Langage de balisage de la variation de séquence génomique*

ISO/HL7 21731:2006, *Informatique de santé — HL7 version 3 — Modèle d'information de référence — Version 1*

CEN EN 13606, *Informatique de santé — Communication du dossier de santé informatisé*

3 Termes et définitions

Pour les besoins du présent document, les termes et définitions suivants s'appliquent.

3.1
acteur
agent
entité qui fournit un stimulus au système

Note 1 à l'article : Les acteurs englobent tant les humains que d'autres entités quasi autonomes, telles que machines, tâches informatiques et systèmes.

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

3.2
allèle
gène trouvé dans différentes formes à la même position dans un chromosome

3.3
BSML
Bioinformatic Sequence Markup Language
spécification de langage extensible et conteneur pour données bioinformatiques

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

3.4
Cell ML
Cell Markup Language
norme permettant de représenter et d'échanger des modèles biologiques informatisés

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

3.5
CGAP
Cancer Gene Anatomy Project
données d'expression génomiques recueillies pour différents tissus tumorigènes chez l'homme et chez la souris

Note 1 à l'article : Le projet CGAP fournit également des informations sur des méthodes et des réactifs utilisés pour obtenir les données génomiques.

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

3.6
codon
séquence de trois nucléotides qui, ensemble, forment une unité de code génétique dans une molécule d'ADN ou d'ARN

3.7
dbSNP
base de données de SNP (4.29) fournie par le National Center for Biotechnology Information (NCBI) des États-Unis d'Amérique

Note 1 à l'article : Disponible sur https://www.ncbi.nlm.nih.gov/SNP/.

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

3.8
DICOM
Digital Imaging and Communications in Medicine
norme dans le domaine de l'informatique médicale pour l'échange d'information numérique entre un équipement d'imagerie médicale (tel qu'une imagerie radiologique) et d'autres systèmes, assurant l'interopérabilité

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

3.9
ADN
acide désoxyribonucléique
molécule qui code l'information génétique dans le noyau des cellules

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

3.10
variation de la séquence d'ADN
différences de séquence d'ADN (4.8) parmi des individus dans une population

Note 1 à l'article : La variation de la séquence d'ADN implique le polymorphisme (4.xx).

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

3.11
DTD
Document Type Definition
document qui contient les définitions formelles de tous les éléments de données dans un type particulier de document HTML (4.15), SGML (4.28) ou XML (4.38)

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

3.12
point d'entrée
point de référence qui indique la (les) classe(s) où les messages débutent pour le domaine

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

3.13
exon
toute partie d'un gène qui encode une partie de l'ARN mature final produit par ce gène après l'élimination des introns par épissage de l'ARN

3.14
médecine génique
médecine basée sur les gènes ou science génétique

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

3.15
GSVML
Genomic Sequence Variation Markup Language
norme pour l'échange de données de la variation de la séquence génomique

[SOURCE : ISO 25720:2009(F)]

3.16
HTML
HyperText Markup Language
ensemble de symboles ou codes de balisage insérés dans un fichier destiné à l'affichage dans un navigateur

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

3.17
CIM-11
Classification Internationale des Maladies, révision 11
outil de diagnostic normalisé pour l'épidémiologie, la gestion de la santé et les applications cliniques

Note 1 à l'article : Disponible sur https://icd.who.int/.

3.18
iCOS
sous-modèle d'informations omiques cliniques pour la CIM

Note 1 à l'article : Modèle d'informations complémentaires pour renforcer la capacité de représentation du modèle de contenu CIM-11 afin de couvrir les informations relatives aux domaines omiques.

3.19
intron
toute séquence nucléotidique dans un gène qui est éliminée par épissage de l'ARN pendant la maturation du produit ARN final

3.20
JPEG
Joint Photographic Experts Group
technique de compression d'images

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

3.21
JSNP
Japanese Single Nucleotide Polymorphisms
base de données japonaise pour les polymorphismes mononucléotidiques

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

3.22
langage de balisage
ML
ensemble de symboles et des règles pour leurs utilisations dans le balisage d'un document

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

3.23
Microarray Gene Expression Markup Language
MAGE-ML
format de données pour décrire des informations sur des expérimentations basées sur le réseau d'ADN et les données d'expression des gènes

3.24
Neuro-ML
Neuro Markup Language
langage de balisage (4.20) pour décrire des modèles de neurones et les réseaux de neurones

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

3.25
nroff
programme de formatage de texte sous systèmes Unix et autres systèmes apparentés à Unix

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

3.26
omique
« omics »
domaine d'étude biologique ayant le suffixe « omique »

Note 1 à l'article : Inclut, sans toutefois s'y limiter, la génomique, la protéomique et la métabolomique.

3.27
pharmacogénomique
branche de la pharmacie visant à développer un moyen rationnel permettant d'optimiser la chimiothérapie, en fonction du génotype du patient

3.28
PolyMAPr
Polymorphism Mining and Annotation Programs
programmes pour l'exploration, l'annotation et l'analyse fonctionnelle de bases données du polymorphisme

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

3.29
polymorphisme
variation de la séquence de l'ADN (4.8) parmi les individus

Note 1 à l'article : Le polymorphisme implique le SNP (4.29) et le STRP (4.32).

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

3.30
ARN
acide ribonucléique
polymère de ribonucléotides se présentant sous la forme de double brin ou de brin simple

[SOURCE : ISO 22174:2005, 3.1.3]

3.31
RNAML
format de données pour l'échange d'informations ARN

3.32
SBML
Systems Biology Markup Language
langage de balisage (4.20) pour les simulations en biologie des systèmes

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

3.33
SGML
Standard Generalized Markup Language
langage de balisage (4.20) pour la représentation de documents qui formalise le balisage et le rend indépendant des systèmes et des traitements

[SOURCE : ISO 8879:1986, 4.305]

3.34
SNP
Single Nucleotide Polymorphism
variation d'un seul nucléotide dans une séquence génétique qui apparaît à une fréquence appréciable dans la population

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

3.35
SNOMED-CT
Systematized Nomenclature of Medicine - Clinical Terms
ensemble dynamique et validé scientifiquement d'infrastructure et de terminologie de soins de santé cliniques

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

3.36
SOAP
Simple Object Access Protocol
protocole léger pour l'échange d'informations dans un environnement réparti décentralisé

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

3.37
STRP
Short Tandem Repeat Polymorphism
segments variables de l'ADN (4.8) qui ont une longueur de deux bases à cinq bases avec de nombreuses séquences répétées

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

3.38
troff
composant principal d'un système de traitement de documents développé par AT&T pour le système d'exploitation Unix

3.39
VNTR
Variable Number of Tandem Repeat
classe de polymorphisme caractérisée par le nombre fortement variable de copies de séquences identiques ou étroitement liées

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

3.40
WML
Wireless Markup Language
langage de XML utilisé pour spécifier le contenu et l'interface utilisateur pour des dispositifs WAP (protocole d'application sans fil)

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

3.41
WGML
Whole Genome Sequence Markup Language
langage de balisage pour représenter la séquence génomique complète

3.42
XHTML
eXtensible HTML
hybride entre HTML (4.5) et XML (4.38) spécialement conçu pour les écrans d'affichage de dispositifs Net

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

3.43
XML
eXtensible Markup Language
version réduite du SGML (4.28), conçue pour les documents Web

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

3.44
schéma XML
langage servant à décrire la structure et contraindre le contenu de documents XML

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

4 Spécifications de l'OML

4.1 Exigences de spécification et positionnement de l'OML (informative)

Le volume des données omiques ne cesse d'augmenter. Ces données sont stockées dans diverses bases de données déclinées en plusieurs formats différents. En outre, les informations d'annotation qui relient les observations du monde réel aux données omiques ne font qu'ajouter au volume et à la complexité des données. Le volume de données, la diversité des structures et la complexité ont rendu difficile l'échange de données omiques. Bien que divers langages de balisage aient été utilisés pour représenter des données omiques, aucun langage centré sur l'omique n'a encore jamais été défini.

L'OML est le premier langage de balisage centré sur l'omique qui soit également orienté vers la santé humaine. Pour être utile dans le domaine des soins de santé humaine, l'OML intègre la capacité d'associer des observations du monde réel à des données omiques. Mais les observations du monde réel peuvent couvrir un large spectre de soins de santé et peuvent être exprimées dans d'autres normes dédiées aux soins de santé. À cet effet, l'OML nécessite une harmonisation avec les autres organismes internationaux de normalisation tels que le Health Level Seven (HL7) et l'Organisation internationale de normalisation [ISO].

Dans le contexte actuel, les informations d'annotation relatives au domaine omique vont en augmentant et ces informations tendent à incorporer les trous d'information. Les données omiques en tant que telles augmentent également mais sont stockées dans différentes bases de données. Le piège dans le traitement des données omiques se situe dans l'absence de normalisation des formats de données pour l'omique organisée. Historiquement, les langages de balisage énumérés ci-dessus ont été utilisés et des programmes sont développés pour gérer l'information omique. Toutefois, il n'existait pas jusqu'ici de langages de balisage centrés sur l'omique. L'OML est le premier langage de balisage centré sur l'omique et axé sur la santé humaine. Considérant que l'omique a un grand impact notamment pour la santé et la réponse humaines, nous pouvons dire que l'OML a le plus grand potentiel d'être le langage de balisage désigné pour la santé humaine. D'une part, la mise en place des applications à la santé humaine dans la pratique signifie qu'il doit gérer les annotations directes ou indirectes. Ici, l'annotation directe indique les informations d'annotation générales telles que l'omique associée à d'autres spécialités omiques et à des préparations expérimentales. L'annotation indirecte indique la totalité des données omiques et données cliniques résultant des données omiques. Pour comprendre la situation clinique omique de chaque patient, nous avons besoin de ces types d'informations supplémentaires. Eu égard à la nécessité d'ajouter de nombreux types d'informations supplémentaires, le développement et la normalisation de l'OML ne peuvent pas être isolés et nécessitent donc une harmonisation avec d'autres organismes internationaux de normalisation tels que le Health Level Seven ou l'ISO.

L'OML est destiné à être utilisé dans les messages d'échange de données liés à la santé humaine. Pour le développement et la normalisation de l'OML dans ce domaine d'application, nous devons toujours garder un œil sur la sécurité du patient, l'efficience clinique et les coûts médicaux. Pour la sécurité du patient du point de vue informationnel, la conservation et la protection des informations relatives au patient sont importantes. Pour le renforcement de l'efficience clinique, la simplicité et l'intelligibilité sans peine sont importantes. Pour la réduction des coûts médicaux, la capacité d'adaptation et la facilité d'installation sont importantes. L'OML tente de satisfaire à ces exigences fondamentales en fournissant le format partageable d'échange de données basé sur XML. L'OML peut être utilisé pour l'échange de données omiques d'un point de vue clinique parmi divers types de formats de données. Dans le cadre plus large de la normalisation des données cliniques, l'OML joue un rôle consistant à décrire les données omiques et leurs informations requises.

4.2 Structure de l'OML (normative)

Le contour de la structure de l'OML est illustré Figure 3. L'OML est constitué de trois critères de données, à savoir les données omiques, les données d'annotation directe et les données d'annotation indirecte. Le critère données omiques décrit, pour chaque domaine omique, les données omiques directes, telles que le type, la position, la longueur, la région, etc. Le critère annotation directe décrit, pour chaque domaine omique, les données connexes des données omiques, telles que l'analyse d'expérimentations, l'épidémiologie ou l'omique associée, etc. Le critère annotation indirecte décrit les informations explicatives/de niveau supérieur des données omiques, telles que les informations cliniques et les données environnementales. Ces critères de données ont intérieurement des relations les uns avec les autres. La structure détaillée de l'OML est illustrée aux Figures 4 à 21.

Une expression OML valide doit être structurée conformément à cette indication.

4.3 DTD de l'OML (informative) et schéma XML (normatif)

Le schéma de l'OML est normatif et disponible sur <lien à définir ultérieurement>. Pour les étapes d'élaboration et de vote, le schéma est fourni dans un fichier OML_Schema_DIS.xsd qui accompagne le présent document.

La DTD de l'OML a été dérivée du schéma ; elle est informative et disponible sur <lien à définir ultérieurement>. Pour les étapes d'élaboration et de vote, la DTD est fournie au format OML_Schema_DIS.xsd et accompagne le présent document.

5 Processus de développement de l'OML (informatif)

Étape 1 : établir les éléments et les besoins selon les cas d'utilisation étudiés à l'aide de l'iCOS CIM-11 de l'OMS.

Étape 2 : construire la structure de base et la DTD.

Étape 3 : étudier le langage de balisage biologique existant, en particulier le GSVML IS25720, et son applicabilité aux besoins (comparaison avec les langages MAGE-ML, BSML, SBML, RNAML [21], ProML, CellML, PolyMAPr)

Étape 4 : affiner la structure de base et la DTD, construire le schéma XML (XSD).

Étape 5 : étudier le format existant (comparaison de leurs formats de données).

Étape 6 : vérifier la capacité d'interface au modèle de génotype du Health Level Seven.