ICS 11.020
CCS C07

T/BIA

团 体 标 准

T/BIA XXXX—202X

生物医学本体
第 1 部分：开发基本原则

Biomedical ontology—
Part 1: Basic principles of development

（征求意见稿）

202X－XX－XX 发布　　202X－XX－XX 实施

北京信息化协会　　发 布

目　次

前　言

本文件按照 GB/T 1.1—2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》的规定起草。

本文件是《生物医学本体》的第 1 部分。《生物医学本体》由以下部分组成：

——第 1 部分：开发基本原则

——第 2 部分：开发规范

——第 3 部分：元数据

请注意本文将的某些内容可能涉及专利。本文将的发布机构不承担识别专利的责任。

本文件由北京信息化协会提出并归口。

本文件起草单位：中国医学科学院基础医学研究所、国家人口健康科学数据中心、中国中医科学院中医信息研究所、北京及遇智悦生物科技有限公司、哈尔滨医科大学、重庆邮电大学、中国医学科学院协和医院、中国医学科学院医学信息研究所。

本文件主要起草人：杨啸林、朱彦、程亮、邵晨、王哲、张敬晨、杨晟、周伟、彭苏元、刘丽红、姚克宇、谢江安、关健、李晓瑛、何勇群、罗葳、张胜发。

引　言

随着生命科学研究范式的演进和生物医学数据的快速增长，多源异构数据的语义集成与智能处理成为推动精准医学与公共卫生治理的关键挑战。本体作为语义建模的知识组织工具，在支撑生物医学信息系统构建与语义互操作方面具有重要作用。

为规范生物医学本体的开发过程，提高本体的可复用性、一致性与互操作性，特制定本系列文件。本系列文件分为以下三个部分：

第 1 部分：开发基本原则。明确生物医学本体在开发过程中应当遵循的基本原则，从方法论层面界定本体开发的总体要求与约束。

第 2 部分：开发规范。规定生物医学本体从规划、设计、实现，到验证与发布的完整开发流程，为本体工程实施提供操作性指导。

第 3 部分：元数据。规定生物医学本体元数据的描述方法，明确核心元数据与扩展元数据的组成及其适用范围。

本文件作为第 1 部分，从方法论层面系统阐明生物医学本体开发应遵循的基本原则，为后续第 2 部分中本体开发流程与实施规范的制定提供理论基础与原则依据。

生物医学本体

第 1 部分：开发基本原则

1 范围

本文件规定了生物医学本体开发所应遵循的基本原则，重点关注本体在语义建模、形式化表示、标识与关系组织、元数据描述以及长期维护与治理等方面的共性要求。

本文件适用于生物医学领域中本体的设计与开发。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本(包括所有的修改单)适用于本文件。

GB/T 41472.2-2022 地理信息 本体 第 2 部分:网络本体 语言(OWL)本体开发规则

GB/T 42986.1-2023 地理信息 本体 第 1 部分:框架

ISO/IEC 21838-1 First edition 2021-08 Information technology Top-level ontologies(TLO) Part 1: Requirements

ISO/IEC 21838-2 First edition 2021-11 Information technology-Top-level ontologies(TLO) Part2: BasicFormalOntology (BFO)

3 术语和定义

下列术语和定义适用于本文件。

3.1

本体 ontology

通过含有定义和公理的基本词汇表对论域现象的形式化表达，这些定义和公理使隐含的意思更明确且可描述现象及其相互关系。

[GB/T 33188.1-2016, 定义 4.1.26]

3.2

顶层本体 top-level ontology

一种旨在代表跨越尽可能广泛领域所共有类别的本体。注：顶层本体有时被称为“形式本体”、“基础本体”、“上层本体”或“领域中立本体”。

[ISO/IEC 21838-1:2021,定义 3.20]

3.3

基本形式化本体 Basic Formal Ontology

基本形式化本体是一种符合顶层本体所规定要求的本体。

3.4

参考本体 reference ontology

一种为特定领域或跨领域知识表示而设计的标准化、形式化的概念框架。它定义了一套共享的、明确的、无歧义的概念，以及这些概念之间的关系，用于描述该领域内的类、属性、关系等。

3.5

统一资源标识符 Uniform Resource Identifier

一种用于标识资源的字符串，它可以是一个资源的地址或一个名字。目的是唯一标识互联网上的资源，无论该资源是可定位的（有地址）还是不可定位的（只有名字）。

[参见 RFC 3986]

3.6

国际化资源标识符 Internationalized Resource Identifiers

国际化资源标识符是统一资源标识符的扩展版本，它允许使用 Unicode 字符，而不仅仅是 ASCII 字符。这使得国际化资源标识符可以包含各种语言和符号（如中文、阿拉伯文、日文等），从而支持全球范围内的语言和字符集。

[参见 RFC 3987]

3.7

语义对齐 semantic alignment

在不同文本、数据或知识表示之间，通过识别并建立它们之间在意义或概念上的对应关系，从而实现信息的匹配、整合和互操作的过程。

3.8

本体复用 ontology reuse

将一个本体或其部分导入到另一个本体中，同时保持所导入内容的意义不变。示例：工具本体的术语被复用到电动工具本体中，后者是前者的一个特化版本。

注释 1：现有本体的术语通常会被复用到新本体中，并与新创建的术语一起出现。

[ISO/IEC 21838-1:2021,定义 3.22]

3.9

外部本体术语最小信息参考 Minimum Information to Reference an External Ontology Term

引用外部本体术语所需的最少信息。

3.10

确定性 certainty

具有明确含义和指向。

3.11

可扩展性 scalability

本体在面临增长或变化的需求时，能够通过增加类、属性、关系、公理增加或改进性能。

3.12

术语 term

在本体论中，术语通常指的是用于表示概念、实体或事物的名称或标签。在本体建模中，术语作为概念的名称，帮助描述和标识类、属性、关系等各种构建块。

3.13

形式语言 formal language

机器可读且具有定义良好的语义的语言

[ISO/IEC 21838-1:2021,定义 3.10]

3.14

公理化 aximination

将一组知识或信息通过形式语言表达为定义和公理集合的过程所产生的结果。

3.15

本体元数据 ontology metadata

对本体本身属性和特征的描述，用于标识、描述、定位、访问和管理本体资源。

注：元数据不仅包括本体的基本信息，如名称、版本、作者、许可协议等，还包括其发布状态、表示语言及访问方式等内容。

4 缩略语

下列缩略语适用于本文件。

BFO 基本形式化本体（Basic Formal Ontology）
GFO 通用形式化本体（General Formal Ontology）
GO 基因本体（Gene Ontology）
LOINC 观测指标标识符逻辑命名与编码系统（Logical Observation Identifiers Names and Codes）
NCIt 美国国家癌症研究所词表（National Cancer Institute Thesaurus）
MeSH 医学主题词表（Medical Subject Headings）
OBO Foundry 开放生物医学本体工厂（Open Biological and Biomedical Ontologies Foundry）
OWL 2 网络本体语言第二版（Web Ontology Language 2）
RO 关系本体（Relation Ontology）
SNOMED CT 医学系统命名法－临床术语（Systematized Nomenclature of Medicine – Clinical Terms）
URI 统一资源标识符（Uniform Resource Identifier）

5 本体开发基本原则

5.1 使用适合的顶层本体

生物医学本体宜采用领域公认的通用顶层本体。在生物医学领域广泛使用的顶层本体有 BFO 和 GFO 等。

5.2 明确领域范围

生物医学本体应在生物医学领域内明确界定其覆盖范围（如临床诊疗、组学研究、药物机制等），并结合应用场景确定粒度与深度，避免概念泛化或跨领域混用。

5.3 遵循生物医学语义一致性原则

生物医学本体在概念体系、术语表达及语义结构上应与现存的主流生物医学知识体系保持一致性与互操作性。生物医学本体的开发应参考主流语义资源（如：MeSH、SNOMED CT、LOINC、NCIt、GO 以及中文临床医学术语等），并结合 OBO Foundry 框架的建模方法和上层逻辑结构（如：BFO、RO），在必要时建立术语映射或交叉引用机制。

对于已有本体或术语系统已覆盖的概念，应优先复用其标识符、语义定义或部分内容；复用的内容可以是已有本体中单个术语或关系，也可以是由多个术语和关系组成的层次结构。对于尚未覆盖的内容，可通过扩展方式提升原有本体内容，例如添加新的细节或更细化的层次结构。语义一致性不等同于框架同一，开发者应以语义互操作为目标，在不同体系间保持概念对齐、标识互参和定义可追溯性。

5.4 使用描述逻辑语言表述

生物医学本体宜采用机器可读、逻辑严谨的形式化语言进行公理化表示，以保证本体结构的清晰性

和语义的明确性，并支持自动化一致性检查与推理。原则上宜采用符合万维网联盟相关规范的描述逻辑型本体表示语言（如：OWL 2 及其后继版本）。

5.5 采用本体标识符

每个生物医学本体应具有唯一的国际化资源标识符，并应为其版本指定相应的版本化标识符，以支持本体资源的发布、维护与升级管理。

5.6 为术语提供准确的定义

在本体中，每一个术语都应该有一个明确的定义，以清晰表达术语的含义。

5.7 关系标准化

生物医学本体开发应使用一套共享的标准关系来定义术语之间的联系。
注：生物医学领域可参考并重用 RO 中定义的标准化关系。

5.8 采用术语和关系唯一标识符

每个术语和关系均应具有一个唯一的标识符。
注：标识符宜采用 URI 进行表示。

5.9 配备本体元数据

生物医学本体应配备完整的本体元数据，并在本体资源注册与管理过程中保持一致和可用。

5.10 建立本体的维持机制

生物医学本体应建立系统化、可持续的维持机制，以支持本体的长期演化与稳定使用。该机制应涵盖版本管理、内容变更控制、术语弃用与演进策略、发布与分发方式、质量保障、维护责任与角色分工，以及用户反馈与社区参与等方面，并明确本体生命周期管理与归档要求。

5.11 遵循隐私与安全保护原则

在生物医学本体的开发与应用过程中，应遵循隐私与数据安全保护的基本要求，防范因语义建模、知识组织或数据标注不当而产生的潜在识别、信息泄露或滥用风险，确保相关活动符合数据安全与伦理合规要求。