



# 中华人民共和国国家标准

GB/T XXXXX—XXXX

## 生物技术 生命科学数据格式和描述要求

Biotechnology — Requirements for data formatting and description in the life sciences

(点击此处添加与国际标准一致性程度的标识)

草案版次选择

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

XXXX – XX – XX 发布

XXXX – XX – XX 实施

国家市场监督管理总局  
国家标准化管理委员会 发布

## 目 次

前言 .....	III
引言 .....	IV
1 范围 .....	5
2 规范性引用文件 .....	5
3 术语和定义 .....	5
4 生命科学数据中对实体和概念描述的建议和要求 .....	11
4.1 概述 .....	11
4.2 推荐的生物和概念实体通用标识符方案 .....	11
4.3 生物实体和概念的数据格式化和上下文描述数据（元数据） .....	13
5 数据格式的技术和组织的建议与要求 .....	13
5.1 概述 .....	14
5.2 组织责任 .....	14
5.3 文档 .....	14
5.4 版本控制和变更日志 .....	14
5.5 兼容性 .....	14
5.6 可扩展性 .....	14
5.7 压缩 .....	15
5.8 结构和控制元素 .....	15
5.9 数据格式中数据类型的要求 .....	15
5.10 一致性和兼容性 .....	16
5.11 数据完整性 .....	16
5.12 格式验证 .....	16
5.13 数据溯源 .....	16
6 数据格式的语义建议与要求 .....	16
6.1 概述 .....	16
6.2 生物数据注释的最小共识 .....	17
6.3 语法和实体化 .....	20
7 适用于生物数据注释的术语和本体的要求 .....	20
7.1 概述 .....	20
7.2 生物本体的要求 .....	20
8 领域特定数据标准的要求 .....	21
8.1 概述 .....	21
8.2 领域特定数据标准的具体要求 .....	21
9 生物数据存储库的要求 .....	22
9.1 概述 .....	22
9.2 生物数据存储库的要求 .....	22

附录 A（资料性） 生命科学数据常见格式示例 ..... 24

    A.1 概述 ..... 24

    A.2 OMICS (组学)、生物化学和分子生物学方法的数据格式 ..... 24

    A.3 生物成像数据的格式 ..... 32

    A.4 应用于生物系统计算机模型的数据格式 ..... 32

    A.5 应用于生命科学模型模拟及其结果的数据格式 ..... 34

    A.6 用于数据和模型质量测量的描述符 ..... 34

附录 B（资料性） 数据、模型和元数据的最低报告标准 ..... 35

    B.1 概述 ..... 35

    B.2 最低报告标准 ..... 35

    B.3 特定领域的本体、分类法和受控词汇表 ..... 37

参考文献 ..... 43

## 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。

本文件由××××提出。

本文件由××××归口。

本文件起草单位：

本文件主要起草人：

## 引 言

生命科学研究及其成果在生物技术中的应用，诊断学和制药业依赖于从广泛的化验、生物学和功能研究中获得的复杂数据，以及过程描述、实验室和现场测量的数据。这包括使用派生数据进行生物、生物技术和生理过程的计算重建、建模和模拟，以及它们在生物技术工作流程中的应用。数据支持的生命科学和生物技术研究跨越了广泛的生物和生物技术领域和应用（例如人类健康、基因工程生物、环境科学、农业、生物修复、DNA测序、色谱、显微镜）。生命科学中的数据驱动、数据密集和大数据分析方法只有使用计算方法并通过数据的一致描述、结构化和集成才可能实现。数据的存储、表示、意义、解释、交换和再利用都受到格式设计的影响。本文件通过为生命科学中的数据记录、处理、重用和交换设定基本要求，满足建立可互操作和明确的数据记录、描述和传输框架的关键需求，从而实现最大的数据价值和利用。

这些来自不同来源、在不同时间记录的生命科学数据必须是可查找、可访问、可互操作和可重复使用的（F-A-I-R）。数据集只有在可访问并以结构良好、一致的格式存储时，才是有价值和有用的。数据版本控制、数据归档和跟踪数据来源由不受时间限制且独立于平台的格式确保。完整且可更新的元数据（即描述数据的数据）有助于数据的定位、使用和分析。

本文件提供了标准化可互操作生命科学数据格式的要求和建议。它为生物技术和生物领域社区定义的许多不同的子领域特定数据格式和描述标准提供了概念框架和参考。为了通俗易懂地利用所引用的特定领域格式化和描述标准及其协同互动而描述了一个最低要求和规则的独立技术框架。因此，本文件提供了相关子领域总体数据格式和通俗的描述规则与指南，作为跨域数据集成的基础。此外，还提供了创建特殊（子）领域的标准、互操作性及其实现的规则和指南。

# 生物技术 生命科学数据格式和描述要求

## 1 范围

本文件规定了生命科学（包括生物技术和生物医学以及非人类生物研究和开发）中数据和相应元数据（即描述数据及其上下文的数据）的一致格式和文档的要求。它为生命科学中的数据呈现提供了可查找、可访问、可互操作和可重用（F-A-I-R）的指导。

本文件适用于为其他目的而系统地捕获、记录或整合生命科学中的数据及其相应的元数据的手动或计算流程。

本文件提供了手动获得的主要实验或程序数据和机器导出数据的要求。本文件还描述了生命科学中数据和相应元数据的存储、共享、访问、互操作性和重用的要求。

本文件规定了从生命科学自动化高通量流程中系统获取大量数据的要求，以及通过其他生命科学技术或手动数据获取的大小规模数据集的要求。

本文件适用于生物技术和生命科学中的许多领域，包括但不限于：生命科学所有领域的基础/应用研究，以及工业、医学、农业、或环境生物技术（不包括用于诊断或治疗目的）及其方法学驱动领域，如基因组学（包括大规模并行测序、宏基因组学、表观基因组学和功能基因组学）转录组学、翻译组学、蛋白质组学、代谢组学、脂质组学、糖组学，酶学，免疫化学，合成生物学，系统生物学，系统医学及相关领域。

## 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。

ISO 8601-1 Date and time — Representations for information interchange — Part 1: Basic rules

ISO 8601-2 Date and time — Representations for information interchange — Part 2: Extensions

## 3 术语和定义

下列术语和定义适用于本文件。

### 3.1

**美国信息交换标准码** American Standard Code for Information Interchange; ASCII

电子通信字符编码标准。

注1：ASCII代码表示计算机、电信设备和其他设备中的文本。

注2：大多数现代字符编码方案都基于ASCII，尽管它们支持许多其他字符。在ASCII文件中，每个字母、数字或特殊字符都用一个7位二进制数（7个0或1的字符串）表示。定义了128个可能的字符。

注3：ISO/IEC 646中记录了7位ASCII。

## 3.2

**向后兼容性 backward compatibility**

较新的编码标准与较老的编码标准的兼容性,其中被设计为以较老的编码标准操作的解码器可以通过对根据较新的编码标准产生的比特流的全部或部分进行解码来继续操作。

## 3.3

**字符 character**

具有语音或象形意义的可打印符号,通常构成文字的一部分,描绘数字或表达语法标点符号。

## 3.4

**特性 characteristic**

限定一个对象或一组对象的属性。

[来源: ISO 1087:2019, 3.2.1, 有修改]

## 3.5

**类 class**

对共享相同属性、操作、方法、关系和语义的一组对象的描述。

## 3.6

**代码 code**

将文本、图像、声音或电、光子或磁信号等信息转换为另一种形式或表示,以便于分析、通信或在存储介质中存储的规则系统。

## 3.7

**概念 concept**

由独特的特征组合创建的知识单元。

[来源: ISO 1087:2019, 3.2.7, 有修改]

## 3.8

**上下文 context**

定义或使用对象的环境、目的和观点。

[来源: ISO/IEC 11179-1:2015, 3.8.7, 有修改]

## 3.9

**数据 data**

以适合通信、解释或处理的形式化方式对信息进行可重新解释的表示。

[来源: ISO/IEC 2382:2015, 2121272, 有修改]

## 3.10

**数据元 data element**

在上下文中被认为是不可分的数据单元。

注1: 该术语是指数据的组织。

注2: 该定义指出,数据元素在某些上下文中是“不可分割的”。这意味着在一个上下文中被认为不可分的数据元素(例如,电话号码)在另一个上下文中(例如,国家代码、区域代码、本地号码)可以是可分的。

[来源: ISO/IEC 15944-1:2011, 3.16, 有修改]

## 3.11

**数据格式 data format**

文件或流中的数据排布。

[来源: ISO/TS 27790:2009, 3.18]

## 3.12

**数据完整性 data integrity**

属性数据没有被未经授权的方式更改或破坏。

[来源：ISO/TS 27790:2009, 3.19]

### 3.13

**数据模型** data model

数据的模型及（或）词汇表示，指定其属性、结构和相互关系。

[来源：ISO/IEC 11179-1:2015, 3.2.7]

### 3.14

**数据提供者** data provider

作为数据来源的个人或组织。

[来源：ISO/IEC/IEEE 5939:2017, 3.5]

### 3.15

**数据集** data set

以一种或多种数据格式可供访问或下载的可识别数据集合。

注1：数据集可以是较小的数据分组，尽管受到某些约束（如空间范围或特征类型）的限制，但在物理上位于较大的数据集内。理论上，数据集可以小到包含在较大数据集内的单个特征或特征属性。

注2：数据集可以以表格形式呈现，并以文字处理文件、电子表格或数据库中的表格形式存储和分发。它也可以以多种替代格式中的任何一种呈现，包括AVRO、JSON、RDF和XML。

[来源：ISO/IEC 11179-7:2019, 3.1.4]

### 3.16

**数据类型** data type

数据分类，表明如何使用数据。

注1：数据类型提供一组值，表达式可以从这些值中取值。

注2：它描述了数据元的内容和结构。

注3：它描述了这些值的属性和对这些值的操作。

注4：数据类型可以以多种方式分类，例如主数据或参考数据。

### 3.17

**数据表示范式** data representation paradigm

数据表示工具，提供明确定义的语法，没有任何应用程序级的语义。

### 3.18

**实体** entity

任何存在、曾经存在或可能存在的特定或抽象事物，包括其属性以及与其他事物的相互作用。

### 3.19

**可扩展性** extensibility

数据格式早期版本中的规定，旨在最大限度地提高该早期版本的实现与该数据格式较新版本的预期实现的互通性。

### 3.20

**向前兼容性** forward compatibility

旧编码标准与新编码标准的兼容性，其中设计用于与新编码标准一起操作的解码器可以解码旧编码标准的比特流。

[来源：ISO/IEC 13818-3:1998, 2.1.108, 有修改]

### 3.21

**标识符** identifier

字符序列，能够在指定上下文内唯一标识与其关联的字符。



[来源: ISO/IEC 11179-1:2015, 3.1.3, 有修改]

### 3.22

**互操作性 interoperability**

两个或者多个系统或组件交换信息并使用已交换信息的能力。

[来源: ISO/TS 27790:2009, 3.39]

### 3.23

**国际化资源标识符 internationalized resource identifier; IRI**

来自通用编码字符集的字符序列,能够在指定上下文内唯一标识与其关联的字符。

注: IRI is an internet protocol element standard that builds on the uniform resource identifier (3.49) by greatly expanding the set of permitted characters.

### 3.24

**JavaScript 对象表示法 JavaScript Object Notation; JSON**

开放且基于文本的交换格式。

注: 以 JSON 格式传输的数据易于读取和写入(对于人类)、解析和生成(对于计算机)。

[来源: ISO/TS 23029:2020, 3.3]

### 3.25

**长期储存 long-term storage**

在一段不确定长度的时间内,永久保留数据。

[来源: ISO 11179:2015, 2.3, 有修改]

### 3.26

**维护员 maintainer**

**维护组织 maintainer**

维护数据格式的个人或组织。

### 3.27

**元数据 metadata**

定义和描述其他数据的数据。

[来源: ISO/IEC 11179-1:2015, 3.2.16]

### 3.28

**元数据对象 metadata object**

由元模型定义的对象类型。

[来源: ISO/IEC 11179-1:2015, 3.2.18]

### 3.29

**元数据属性 metadata attribute**

其规范中通常需要的元数据对象实例的属性。

### 3.30

**命名空间 namespace**

用于标识和引用可实例化为统一资源标识符的各种对象的元素类。

注1: 命名空间确保所有给定的对象集都具有唯一的名称,以便识别。

注2: 命名空间通常被构造为层次结构,以允许在不同上下文中重用名称。

### 3.31

**对象 object**

任何可感知或可想象的事物。

注: 对象可以是物质的(例如“发动机”、“纸张”、“钻石”)、非物质的(例如“转化率”、“项目计划”)

或想象的（例如“独角兽”、“科学假设”）。

[来源：ISO 1087:2019, 3.1.1]

### 3.32

**本体论** ontology

术语、关系表达式和相关自然语言定义的集合，以及一个或多个旨在捕捉这些定义的预期解释的形式理论。

注：本体定义了一组表征原语，用于对知识或话语领域进行建模。表示原语通常是类（或集合）、属性（或特性）和关系（或类成员之间的关系）。表示原语的定义包括关于其含义的信息和对其逻辑上一致的应用的约束。

[来源：ISO 23903:2021, 3.18]

### 3.33

**网络本体语言** web ontology language; OWL

基于Web的语言，设计用于需要处理信息内容的应用程序。

[来源：ISO 14199:2015, 3.6]

### 3.34

**容许值** permissible value

价值含义的指定。

[来源：ISO/IEC 11179-1:2015, 3.3.20, 有修改]

### 3.35

**永久标识符** persistent identifier; PID

唯一标识符，通过提供独立于其物理位置或当前所有权的访问来确保对数字对象的永久访问。

[来源：ISO 24619:2011, 3.2.4, 有修改]

### 3.36

**谓词** predicate

**限定符** qualifier

数据集或数据元素与引用资源的特定主题之间的关系。

### 3.37

**属性** Property

类所有成员共有的特征。

### 3.38

**专有软件** proprietary software

非免费的计算机软件，其软件发布者或其他人保留其知识产权，通常是源代码的版权，有时也包括专利权。

### 3.39

**出处** provenance

有关资源的起源、派生或生成的地点和时间的信息或真实性或过去所有权的记录或证明。

[来源：ISO/IEC 11179-7:2019, 3.1.10]

### 3.40

**出版商** publisher

已发布数据格式的个人或组织。

### 3.41

**数量** quantity

**定量值** quantitative value

现象、物体或物质的特性，该特性具一定规模，能以可用的计数表示，且具有参考性。

[来源：ISO/IEC Guide 99:2007, 1.1, 有修改]

### 3.42

**资源描述框架** Resource Description Framework; RDF

用于描述元数据的XML语法。

[来源：ISO 16684-1:2019, 3.6]

### 3.43

**具体化** reification

使一个主题代表同一主题地图中另一个主题地图构造的主题。

[来源：ISO/IEC 13250-2:2006, 3.11]

### 3.44

**仓库** repository

**数据仓库** data repository

一系列数据的呈现，以及相应的数据访问和控制机制，如搜索、索引、存储、检索和安全性。

注：存储库可以涵盖数据治理、数据管理和数据所有权的各个方面。

[来源：ISO/IEC 20944-1:2013, 3.21.12.19, 有修改]

### 3.45

**语义互操作性** semantic interoperability

系统共享的数据在正式定义的领域概念层面上被理解的能力。

[来源：ISO/TS 27790:2009, 3.67]

### 3.46

**稳定格式** stable format

**稳定的数据格式** stable data format

数据格式规范不会随时间发生持续或重大变化。

### 3.47

**术语** term

通过语言手段表示一般概念的名称。

[来源：ISO 1087:2019, 3.4.2, 有修改]

### 3.48

**专业术语** terminology

表示指定领域内的概念的术语集合。

注：这意味着已发布的目的和范围，从中可以确定该表示形式充分覆盖指定领域的程度。

[来源：ISO 1087:2019, 3.1.11, 有修改]

### 3.49

**统一资源标识符** uniform resource identifier; URI

用于唯一标识抽象或物理资源的紧凑字符序列。

注：请参阅 IETF RFC 3986:2005。

[来源：ISO/IEC 12785-1:2009, 3.23, 有修改]

### 3.50

**计量单位** unit of measure

测量相关值的实际单位。

注：相关概念域的维度必须适合指定的测量单位。

[来源：ISO/IEC 11179-1:2015, 3.3.29]

## 3.51

**通用编码字符集** universal coded character set; UCS  
国际电子通信字符集编码标准。

## 3.52

**验证** verification

通过提供客观证据确认特定要求已得到满足。

注：验证所需的客观证据可以是检查的结果或其他形式的确定结果，例如执行替代计算或审查文件。

[来源：ISO 9000:2015, 3.8.12, 有修改]

## 3.53

**可扩展标记语言** extensible markup language; XML

以机器可处理且人类可读的方式对信息进行编码的标记语言。

[来源：ISO 5127:2017, 3.1.9.19]

## 4 生命科学数据中对实体和概念描述的建议和要求

## 4.1 概述

本节重点关注生命科学数据和数据类型（参见 ISO/IEC 11404）中生物或概念实体的一致描述的建议和要求，以及使用普遍持久标识符（PIDs）对其进行明确引用。

数据集、相应的元数据集、或数据集中的任何生物或概念实体或定义的过程，都应该是明确可识别的。因此，应使用统一资源标识符（URIs）或国际化资源标识符（IRIs），来赋予生物或概念实体或定义的过程相应的明确定义或实体的引用或过程。通过使用相应的URI或IRI对数据集中的实体或过程进行注释，可以实现以上目标，这些实体来源于数据库、注册表、术语资源、本体或其他合适的资源，其中合适的资源包括携带相应的定义或用于消除实体或过程中歧义的数据条目。

## 4.2 推荐的生物和概念实体通用标识符方案

## 4.2.1 URI 规定

## 4.2.1.1 概述

数据集中生物或概念实体或定义的过程，以及相应的元数据集或数据集中，可以表示为或注释为URI。如果生物或概念实体标识符具有特定的命名空间和上下文（例如，留存在数据库或包含在特定引用中）。生物或概念实体的URI，可以用任何适当的可兼容方式表示，例如http、https、urn等。尽管非注册也仍然有效，但使用的URI方案应在互联网编号分配机构（IANA）进行注册。URI应为ASCII字符串，格式如下：

scheme://authority/path/name

其中“authority”和“path”定义了数据的类型（命名空间），即同一类型的所有“name”的集合，“name”指的是该命名空间内各自的生物或概念实体。“authority”至少应包括主机，由命名空间所指的注册名称或IP地址组成（例如承载命名空间或指向它的数据库或Web资源），“path”包括指向相同类型的名称集合的至少一个或多个分层结构的名称空间限定符。路径内的层次级别应由正斜杠（“/”）定义。在ASCII字符集中，字符： / ? # [ ] @ 保留用作通用URI组件的分隔符，并且应进行百分号编码（“转义”），例如“%3F”代表问号。

对URI进行解引用应指向对不同生物实体或概念的表示，这些实体或概念应由URI标识的。如果两个URI的转义版本在每个字符上均相同，则被认为是相同的。不同的URI可以等效，但必须由软件代理进行规范化处理。

#### 4.2.1.2 URI 的持久性

任何用于描述数据或其中任何实体（或两者）的URI都应该是持久且不可更改的。对于由URI标示的数据的更改，应该保留其来源和版本信息。

#### 4.2.1.3 URI 的元数据

与URI相关联的任何元数据都应该是可捕获的，并在数据的整个生命周期内保留。

URI应该是持久的，并且与其在服务器上的映射和表示形式（包括大小写字母）无关。尽管方案不区分大小写，但对于指定方案的文档，规范形式应为小写。为了稳健性，实现可以接受大写字母作为方案名称的等效形式（例如，允许使用“HTTP”和“http”）。URI不应包含可推断生物实体或概念属性的信息。

URI应仅标识一个生物实体或概念。使用同一URI来标识多个生物实体或概念会导致URI冲突，应避免URI冲突。数据库社区有责任避免将等效URI分配给多个生物实体或概念。数据库社区负责URI的表示管理。

URI 应是不透明的，并且不应包含：

- a) 作者姓名；
- b) 主题；
- c) 状态；
- d) 路径；
- e) 文件扩展名；
- f) 软件机制；
- g) 磁盘名称；
- h) 域名。

#### 4.2.2 IRI 规定

数据集中的生物或概念实体或定义的过程、相应的元数据集或数据集合可以表示为 IRI 或由 IRI 注释。IRI 是 URI 的补充。它扩展了URI的语法，适用于更广泛的字符集，并定义了与其他结构（如URI引用）相对应的“国际化”版本。

IRI 应用于所有不只使用 ASCII 字符的实体。4.2.1中的所有其他 URI 规定也相应地适用于IRI。

IRI 应表示为来自 UCS 的字符序列，即通用编码字符集（符合 ISO/IEC 10646 的 Unicode）。IRIs 应是来自 UCS 的字符串，格式如下：

scheme://authority/path/name

其中“authority”和“path”定义了数据的类型（命名空间），即同一类型的所有“name”的集合，“name”指的是该命名空间内各自的生物或概念实体。路径内的层次级别应由正斜杠定义。

#### 4.2.3 URI 和 IRI 之间的关系

尽管ASCII字符集包含在UCS中，且可读，但反之不成立。因此，可能需要执行映射的步骤，以将URIs检索对应到IRIs。

IRIs 应需要一个转换步骤以方便转换为 URI。

### 4.3 生物实体和概念的数据格式化和上下文描述数据（元数据）

#### 4.3.1 概述

生物实体和概念的数据和元数据格式可能因社区、学科、机构、国籍和时间而异。本文档确保在不同社区、学科、机构和国籍间，数据和元数据的可用性在时间尺度上受到最小的影响（参见 ISO/IEC 14957 和 ISO/IEC TR 10032）。

#### 4.3.2 版本控制

包括生物实体的数据和元数据格式以及基础概念的所有方面都应进行版本控制。使用版本控制系统可以实现对生物实体和概念的版本控制。数据格式的元素应使用已建立的本体进行版本控制（见附录B）。

#### 4.3.3 任意限制

生物实体和概念的格式不应包含任意限制，例如字符串的最大长度或强制使用或区分大小写限制。

#### 4.3.4 字符集

生物实体和概念的格式应支持所有语言的 ASCII 和非 ASCII（即 UCS）字符。

#### 4.3.5 机器可读性

用于描述和编码生物实体和概念以及其相应的元数据的格式应确保机器可读性，并也需保证人类可读性（参见ISO/TR 3985）。

若基于手动生成的纯文本，创建用于描述和编码生物实体和概念以及其相应元数据的格式，需防止其受到破坏性的更改。

为描述和编码生物实体和概念以及其相应的元数据，而构建数据和元数据格式时，需使用广泛认可的数据表示范式，例如JavaScript对象表示法（JSON）、可扩展标记语言（XML）和资源描述框架（RDF）或类似概念，合适的领域特有的元数据标准和格式，以及公认的术语（见附录B中推荐的元数据标准和领域特定术语）。对于一致的数据表示和结构化，只有在适用的情况下才应使用公认的领域特定数据格式（见附录A）和元数据模型（参见参考[10]和ISO/IEC 19502）。

不具有开放性并且在处理或传输过程中无法保护语义上下文丢失的数据和元数据格式（即在数据库中）应使其具备机器可读性，且应符合以下条件：

- a) 安全考虑；
- b) 成本和收益；
- c) 法律责任；
- d) 知识产权；
- e) 商业机密；
- f) 合同限制；
- g) 其他具有约束力的书面协议。

#### 4.3.6 知识表示

知识表示应使用本体创作框架，例如Web本体语言（OWL）或类似范例，包括 JSON、XML和 RDF。可以利用推理器来确保逻辑一致性。

## 5 数据格式的技术和组织的建议与要求

## 5.1 概述

数据格式可以通过不同的方式结构化。这些结构取决于生成数据的过程、数据的预期用途以及正确解释数据所需的元数据量，并使其易于寻找、可访问、可互操作和可复用（“F-A-I-R”）。

## 5.2 组织责任

负责建立、维护和/或潜在更改数据格式的组织应被记录。每个组织应提供相应的联系信息（例如电子邮件地址、网站）。

有关数据格式的信息应至少包括：

- a) 格式描述；
- b) 版本；
- c) 结构；
- d) 数据表示。

数据格式的维护员应对以下事项负责：

- 用户请求；
- 格式更新；
- 规范中的错误更正。

数据格式中的所有实体、属性、过程和特征的数据表示和格式应保持一致。

## 5.3 文档

应提供以下内容的全面文档：

- a) 来源；
- b) 维护；
- c) 格式结构；
- d) 数据项；
- e) 数据格式化；
- f) 格式的特性。

以上属性任何一项都应以适当方式的提供。应提供一个稳定且可识别的来源，包含有关数据格式的信息，并进行维护和更新。

应记录数据类型及其元数据的表示方式。

注：文档可以以电子文件、在线或作为扫描的打印文档（最好是PDF格式）的形式提供。

## 5.4 版本控制和变更日志

数据格式应包含确切格式版本的信息（如适用，则也需包含子版本）。对于结构简单且不太可能发生格式变更的稳定格式（例如FASTA），可豁免此规则。在这种情况下，应提供关于豁免的注释信息。

对格式进行的更改应进行文档记录，并通过更改版本号进行指示。描述格式的元数据应与其包含的数据相关联。

## 5.5 兼容性

应确保对先前版本的格式进行向前兼容。如适用，应提供与先前版本的向后兼容性。

## 5.6 可扩展性

应确保在不影响兼容性的情况下，为格式的未来版本添加新的数据项。

## 5.7 压缩

对于压缩的数据记录，应引用压缩算法或适当的压缩和解压缩工具。对于自定义的压缩技术，数据提供者或维护员应确保压缩和解压缩工具的完整性。在这种情况下，应披露完整的压缩/解压缩算法。解压缩应保持原始数据的完整性。

## 5.8 结构和控制元素

应对具有特殊含义的元素（例如字段或记录分隔符、转义序列、换行符或类似元素）进行文档记录。

## 5.9 数据格式中数据类型的要求

### 5.9.1 概述

在适用的情况下，数据类型的表示应遵循公认的标准（IEEE、ISO等）。

### 5.9.2 数值编码

应对数值型数量值的表示进行文档记录。如果适用，应使用标准的格式（如IEEE 754）。如果数值型数量值由字符串表示，应指定十进制分隔符、指数表示以及前缀。

对于非标准表示，应指明相应格式中允许的值范围。非十进制数据应明确指出（并相应地记录规范）。如果适用，数值型数量值应标注为测量、推断或假设数据。

测量数据应包括以下信息：

- a) 测量精度；
- b) 测量准确度；
- c) 测量不确定性；
- d) 如有记录，则应提供获取数据的方法。

适用的数量/数量数据应分配以国际单位制（SI单位）表示的适当计量单位。如果单位无法用SI系统表示，应提供适当的换算因子。有序数值数据应指定为有序数值，并指定适当的范围或允许的值。

### 5.9.3 字符串的编码

对于由字符串表示的数据（可读数据），应指定编码（例如ISO/IEC 8859系列、Unicode），除非按照ISO/IEC 646（ASCII）使用编码。

### 5.9.4 测序编码

核酸和蛋白质序列数据应根据国际纯粹和应用化学联合会（IUPAC）和国际生物化学和分子生物学联合会（IUBMB）在“生化命名和相关文件”（即白皮书）中的建议进行编码，该建议由IUPAC-IUBMB生化命名联合委员会和IUBMB命名委员会发布。

### 5.9.5 时间

日期和时间的呈现应符合ISO 8601-1和ISO 8601-2中指定的格式。任意时间数据（如时间序列数据）可以按照ISO 8601-1和ISO 8601-2表示，也可以按照5.9.2中指定的测量数据表示。

### 5.9.6 布尔数据

应指定布尔状态和允许的值的表示方式。如适用，应使用0或1来表示布尔状态，其中0表示“假”，1表示“真”。



### 5.9.7 生物成像数据

图像应以公认的标准格式进行编码，最好以原始数据的形式使用无损压缩格式保存所有成像数据，例如TIFF（标签图像文件格式，参见ISO 12639）。如果需要缩小图像大小，也可以使用已建立的图像压缩格式，如JPEG（参见ISO/IEC 10918-1）或PNG（参见ISO/IEC 15948）。如果可转换为标准格式，则可以使用专有图像格式（例如来自显微镜制造商）。

对于来自健康和生物医学领域的图像数据，应考虑使用数字影像和医学通讯标准（DICOM）（参见ISO 12052）或类似广泛使用的格式。

### 5.10 一致性和兼容性

应通过适当的验证手段确保数据表示的内部一致性和兼容性，数据元素之间的关系（及其相应的元数据），术语和使用的词汇表。此外，应通过适当的验证手段确保格式与生命科学或技术领域中有格式之间的结构和语义具备互操作性。应检查格式与具有交叉或潜在连接点的技术领域的格式之间的互操作性。

在格式转换过程中不应丢失数据和语义上下文。然而，如果发生数据和语义上下文的准确性损失，应进行记录。应通过适当的验证手段确保数据表示的内部一致性和完整性，数据元素之间的关系（及其相应的元数据），术语和词汇表。应验证格式与相同生命科学技术领域的现有格式以及与具有交点或潜在接触点的技术领域的格式之间的结构和语义互操作性。

格式的数据结构应根据相应领域适用的定义进行开发（示例请参见附录B.2），且应符合最低报告准则。通过对领域特定本体论、分类法和受控词汇进行注释，使用对应于解析实体定义的相应资源的唯一PID，应定义所包含项、过程和特征的上下文和生物学含义（示例请参见附录B.3）。

### 5.11 数据完整性

数据格式应保证格式和数据完整性的可检查性，例如使用校验或其他类似的方法。

### 5.12 格式验证

格式的发布者或维护员应使得数据格式的强度、弱点、适用性和限制可验证，并提供一种使数据文件可以根据格式规范进行验证的方法。

### 5.13 数据溯源

应使用结构化、可互操作且可机器处理的溯源信息来记录数据的完整历史。通过连接描述任何先前处理步骤、方法、工具、生物实体、生物材料和用于生成所记录数据的元数据，应保持连续的溯源信息链。完整的溯源信息链既可以评估数据质量和其特定的目的，也可以通过追踪其起源、生成、处理和分析来建立数据的可靠性和可重复性。如果溯源信息包含敏感或个人可识别的信息，应采取适当的预防措施，如访问控制机制。为了在异构环境（如Web）中实现可互操作的溯源信息交换，应为每种格式的所有数据建立一个定义的模型、相应的序列化和其他支持定义（例如，W3C PROV标准）。

## 6 数据格式的语义建议与要求

### 6.1 概述

格式应具有明确的结构，应有利于数据的提取和元数据的可用性。对于相同数据类型的所有项、过程和特征，数据的表示和格式应保持一致和统一。应记录数据类型及其元数据的表示方式。在适用的情况下，数据类型的表示应遵循被普遍接受的标准（如IEEE、ISO等）。

定位数据的能力取决于一组明确定义的生物描述符。应以一致的方法表示生物实体的描述，以便相应的搜索、分析和挖掘工具可以在整个生命科学数据领域范围内定位到描述性数据。本标准对开发领域独立的生物数据注释形式进行了要求的定义。这些生物描述符在生命科学数据领域之间，应兼容于语法和语义，以确保数据共享、保持、访问、使用和可重复使用。

实体应以相关生物本体论提供的最具体的术语作为URI进行注释。如果没有精确的术语，则应使用最接近的常规术语。如果没有足够精确的术语，注释者应请求相关生物本体论维护组添加更精确的术语。此外，术语的可读版本应作为注释包含在内。

如果一个概念在多个本体论中出现，则应使用与数据集最相关的本体论。例如，组织类型通常在多个本体论中定义，应使用最接近的可用的物种特定的本体论。

## 6.2 生物数据注释的最小共识

### 6.2.1 概述

最低要求的注释，应描述数据集的上下文语义、参数和结果，包括生物、医学和环境的背景信息。这些元数据注释应简洁地描述产生数据集的过程（例如分析或实验程序）、基本目标（例如解决的问题）、相应的上下文、组件、独立（受控、变化）量和依赖的可测量量。这些注释可通过简明的表格表示；可以转换为标准的语义格式，例如RDF；或者编写为能够与W3C数据相关的“链接开放数据”概念进行兼容。注释的语法是一系列的三元短语，形式为主语 - 谓语 - 宾语。例如：“肝脏” - “是一个” - “器官”。表2列出了应用于数据注释的谓语的示例。精确的语法和实体化应“与目的相符合”。例如，对于使用标准网络搜索引擎进行数据搜索和检索，语法应符合这些搜索引擎的索引。

下面描述了注释中需要包含的必需项和建议项（另见表1，其中列出了基本的必需生物描述符的示例）。此外，还给出了每个项所建议使用的谓语。例如，应使用指向相应术语或条目的URI对肝细胞进行注释，该URI指向引用资源（例如受控词汇、领域本体论或术语），并使用有解析服务的URI，以保证持久可解析性（例如，参考文献[19]中引用的URI指向和引用了解剖学基础模型本体论（FMA）中相应术语“肝细胞”的URI）。此外，为了便于阅读，还可以用相应的通用名称，例如“肝细胞”、“肝实质细胞”等对其进行注释。

注：尽管RDF框架被用作描述框架的示例，但其他框架（例如JSON、XML）也可用于描述语义信息。

### 6.2.2 物种

注释应包括所研究、处理和分析的物种的描述。在物种的指定不够精确的情况下，如果有记载，应使用品系、品种或其他更精确的术语。例如，在细菌研究或微生物生物技术分析中，除了细菌名称外，还应使用细菌菌株、血清型或分子亚型。所使用的谓语应为“是”。如果注释无法准确识别正确的（亚）物种，应参考更高级别的分类术语，则应使用谓语“是...的样式”，例如“是哺乳动物的样式（方式）”。

### 6.2.3 性别

如果适用，注释应包括被分析或研究个体的性别描述。在性别不适用的情况下（例如在细菌中），此项应为“不适用”。如果性别适用但未知，则此项应为“未知”。所使用的谓语应为“是”（参见ISO/IEC 5218）。

### 6.2.4 年龄

如果适用，注释应包括被分析或研究个体的年龄（或年龄范围）描述。在年龄不适用的情况下（例如在细菌中），此项应为“不适用”。如果年龄适用但未知，则此项应为“未知”。所使用的谓语应为“是”。

### 6.2.5 器官

如果适用，注释应包括被分析或研究的器官的描述。在器官指定不适用的情况下（例如在细菌或其他微生物中），此项应为“不适用”。如果器官适用但未知，则此项应为“未知”。所使用的谓语应为“是”。

#### 6.2.6 组织

如果适用，注释应包括被分析或研究的组织的描述。在组织指定不适用的情况下（例如在细菌或其他微生物中），此项应为“不适用”。如果组织适用但未知，则此项应为“未知”。所使用的谓语应为“是”。对于复杂的组织结构，或者如果注释无法准确识别正确的组织，且参考和引用了本体论中的更高级别术语，则应使用谓词“是...的一部分”，例如“是血液造血系统的一部分”。

#### 6.2.7 细胞类型

如果适用，注释应包括被分析或研究的细胞或细胞群的描述。在细菌或其他微生物中不适用细胞或细胞群的情况下，此项应为“不适用”。如果适用但未知，则此项应为“未知”。所使用的谓语应为“是”。如果注释无法准确识别正确的细胞类型，并因此参考了引用本体论中的更高级别术语，则应使用谓词“是...的样式”，例如“白细胞是造血细胞的样式”。

#### 6.2.8 可识别对象

注释应包括分析工作流程或研究中相关可识别对象的描述。在分析工作流程或实验中，可识别对象是指在过程中可以看到和/或测量的任何有形对象。分析工作流程或实验中可以有多个可识别对象。所使用的谓语应为“是”。

#### 6.2.9 可识别过程

注释应包括在产生数据的分析工作流程或实验中的可识别过程的描述。可识别过程是指分析工作流程或实验中随时间变化且可测量的任何组成部分，例如细胞增殖或死亡。一个实验中可以有多个可识别的过程。所使用的谓语应为“是”。如果注释无法准确识别正确的过程，并因此引用本体论中的更高级别术语，则应使用谓词“是...的样式”。

#### 6.2.10 操作实体

注释应包括在产生数据的分析工作流程或实验中操作实体的描述。操作实体是指分析工作流程或实验中被实验者改变和控制的任何组成部分，例如向细胞培养基中添加生长因子或根据不同的群体成员特征进行分组。一个实验中可以有多个操作实体。所使用的谓语应为“是”。

#### 6.2.11 分析、实验和计算

注释应包括用于执行注释过程的分析、实验和/或计算技术的描述。所使用的谓语应为“是...的样式”。

#### 6.2.12 生物学问题或分析问题

对于分析或实验过程，注释应包括分析工作流程或实验所设计解决的生物学问题的描述。应描述一个高级别的生物过程，例如疾病状态、正常稳态控制过程、发育过程等。所使用的谓语应为“是”。

#### 6.2.13 技术特定数据

数据本身应使用所应用技术的相关领域特定标准格式进行编码（参见附录A中的推荐示例）。数据的描述应使用适当的领域特定标准元数据文档格式进行编码（参见附录B中的推荐示例），遵循相关的最小信息标准，提供元数据对象及其相应元数据属性的检查清单，以便在特定领域和/或特定分析工作

流程或实验设置中进行文档化（参见附录B. 2中的推荐示例），并使用适当的领域特定本体论、分类法和受控词汇表（参见附录B. 3中的推荐示例）进行数据的规范和注释。

表1中显示了基本所需的生物学描述符的示例。

表 1 基本所需的生物学描述符示例

字段名称 (主体)	谓语	建议的本体论或 词汇表	注释	对象 - 人类可读示 例 <sup>a</sup>
物种	是	NCBI taxonomy	实验所涉及的物种。	人类，大肠杆菌
性别	是		测试对象的性别，或者适用的情 况下，来源组织或细胞的性别。	男性，女性，男-女， 女-男，两性同体，以 及其他适用的选项
年龄	是		进行研究的个体的年龄，或提供 样本的个体的年龄。	岁，受精后 8 小时 (HPF)
器官	是	FMA20	样本的器官来源。	肝脏，不适用
组织	是	FMA20	样本的组织来源。	实质组织，不适用
细胞	是	FMA, Uberon	样本中可识别和/或可观察到的 细胞类型。在细菌和其他单细胞 生物的情况下，可以重复填写。	肝细胞
可识别对象	是	protein, GO	实验中的任何测量数量。包括因 变量和自变量。(因变量)	细胞计数增加 (因变 量)，基因缺失
可识别过程	是	GO	在实验中可以直接观察到的过 程。(因变量)	细胞增殖，细胞死亡， 细胞分裂，小分子代 谢
操作的实 体	是	蛋白质，GO，小 分子，环境	在实验中变化的实验性特性。 (自变量)	添加 IL-1，改变营养 浓度
实验技术	是...的样 式	GO, OBI	实验中使用的技术。	微阵列，细胞培养， 显微镜图像
生物学问 题	是...的样 式	GO, NCI 词库	实验旨在回答的基本生物过程 或生物学问题。	细胞增殖的刺激，毒 物效应，胚胎发育

<sup>a</sup> 实际数据还应包括指向特定本体的 URI。

谓语的示例在表2中显示。

注：基于COMBINE/BioModels.net限定符。

表 2 谓语（限定符）示例

谓语	描述
是	数据集元素所表示的生物实体或过程与引用资源的主题具有相同的身份。该谓语用于将数据集的组成部分与另一个资源、受控词汇或本体论中的确切表示相连；例如，将数据集中的肝细胞与本体中的“肝细胞”术语相连接。
是被...描述	数据集元素所代表的生物实体或过程由所引用资源的主题描述。例如，可以使用这种关系将物种或参数与描述该物种浓度或该参数值的文献联系起来。
具有...的部分	数据集元素所代表的生物实体或过程包括所引用资源的主题，无论是在物理上还是逻辑上。例如，可以使用这种关系将细胞与其包含的亚细胞部分相连，或者将

	多组分蛋白质复合物的组分描述相连。
是…的一部分	数据集元素所代表的生物实体或过程是所引用资源主题的物理或逻辑部分。可以使用这种关系将数据集组件与包含它的复合物描述相连。例如，可以使用这种关系将亚细胞部分与包含它的细胞相连，或将多组分蛋白质复合物的组分描述与复合物本身相连。
是…的样式	数据集元素所代表的生物实体或过程是所引用资源主题的一个版本或实例。这种关系可以用来表示特定生物实体的“超类”或“父类”，例如。
有…的形式	所引用资源的主题是数据集元素所代表的生物实体或过程的一个版本或实例。这种关系可以用来表示生物实体的异构体或修改形式，例如，一个酶类的同工酶。
编码	数据集元素所代表的生物实体或过程直接或间接地编码所引用资源的主题。这种关系可以用来表达，例如，特定的 DNA 序列编码了一种特定的蛋白质。
有…的属性	所引用资源的主题是数据集元素所代表的生物实体或过程的一个属性。这种关系可以在生物实体展示某种酶活性或发挥特定功能时使用。
是被…编码	数据集元素所代表的生物实体或过程直接或间接地由所引用资源的主题编码。这种关系可以用来表达，例如，一种蛋白质由特定的 DNA 序列编码。
是与…同源	数据集元素所代表的生物实体或过程与所引用资源的主题具有同源关系。这种关系可以用来表示共享共同祖先的生物实体。
发生于…	数据集元素所代表的生物实体或过程在物理上限制在一个位置，该位置是所引用资源的主题。这种关系可以用于描述一个反应发生的区域位置，或者描述一个过程发生的生物体或生物体部分。
具有…粉类群	数据集元素所代表的生物实体在分类上受到限制，而该限制是所引用资源的主题。这种关系可以用于将物种限制归属于生化反应。
是…的属性	数据集元素所代表的生物实体或过程是所引用资源的一个属性。

6.3 语法和实体化

数据可以以表格形式表示（如表1），也可以存储在数据库或其他数据资源中。无论哪种情况，注释都应该能够被实体化为RDF三元组和纯文本。在所有情况下，实体化都应该被证明是“适合目的”的。

7 适用于生物数据注释的术语和本体的要求

7.1 概述

用于描述生命科学领域中的数据、概念和数据实体的术语和本体应该有助于识别和理解本体所涵盖的生物或生物技术领域的关键概念。

7.2 生物本体的要求

7.2.1 维护员

本体应该有一个明确定义的维护员，与相关的社区保持一致。该社区应对任何对此感兴趣的个人或组织开放。该组织应该有一个网络存在。

7.2.2 本体的维护

本体的维护员应该有一套明确定义的维护程序。其中确保包含维护网络，同时确保本体符合其领域和使用的要求。

### 7.2.3 本体语法

维护员应根据社区的需求定义本体的语法。语法应基于现有的本体基础设施，如W3C OWL或其他广泛使用的语法，如开源的生物和生物医学本体论（OBO）工坊。

### 7.2.4 与其他本体的链接和术语复用

维护员应尽可能地与其他相关本体进行链接。

### 7.2.5 授权和归属

维护员应发布本体的授权模型。

维护员应发布本体的归属模型。

### 7.2.6 稳定的 URI 和版本信息

维护员应提供一种机制，用于创建和共享本体中术语和概念的稳定 URI。

维护员应提供一种机制，用于对个别术语和整个本体的版本进行版本控制。

### 7.2.7 社区

维护员应将受本体影响最大的社区参与到创建、扩展和维护本体的过程中。

### 7.2.8 语言

本体中使用的人类可读语言应由维护员决定。

维护员应根据需要和实际情况尽可能地使本体具备多语言属性。对于国际范围的使用，控制词汇表和本体应提供英文的版本。

## 8 领域特定数据标准的要求

### 8.1 概述

“领域特定数据标准”可以指应用方法定义的技术领域（显微镜、微阵列等），也可以指基础过程定义的生物领域（癌症、生殖等）。

### 8.2 领域特定数据标准的具体要求

#### 8.2.1 维护员

数据标准应有一个明确定义的维护员，与相关社区保持一致。该社区应对任何对此感兴趣的个人或组织开放。该组织应该有网络相互连接。

#### 8.2.2 数据标准的维护

数据标准的维护员应有一套明确定义的维护程序。确保有维护用的网络，以及符合其领域和使用要求的标准。

应定义流程，进行数据标准中术语或结构的添加、删除和/或废弃。

#### 8.2.3 数据标准的语法

维护员应根据社区的需求定义数据标准的语法。语法应基于现有的框架，如XML、JSON和RDF。

#### 8.2.4 与其他数据标准的联系

维护员应：

- a) 尽可能与其他相关数据标准进行链接；
- b) 在适当的情况下复用其他领域已有的数据标准；
- c) 使用符合标准的生物本体中的术语。

数据标准应与符合标准的数据存储库兼容。

#### 8.2.5 授权和归属

维护员应发布：

- a) 数据标准的授权模型；
- b) 数据标准的归属模型。

#### 8.2.6 稳定的 URI 和版本信息

维护员应提供一种机制：

- a) 用于创建和共享术语、概念和整个数据标准的稳定 URI；
- b) 用于对数据标准进行版本控制。

#### 8.2.7 社区

维护员应将受数据标准影响最大的社区纳入到创建、扩展和维护数据标准的过程中。

#### 8.2.8 语言

数据标准中使用的人类可阅读语言应由维护员决定。

维护员应尽量使数据标准在需要的情况下具备多语言的版本。对于国际范围的使用，应提供控制词汇表和本体的英文版本。

### 9 生物数据存储库的要求

#### 9.1 概述

生物数据存储库应推动生物数据的长期存储，包括存档、索引、搜索和共享。数据存储库中使用的数据格式应符合本文件的要求。

#### 9.2 生物数据存储库的要求

##### 9.2.1 维护员

存储库应有一个明确定义的维护员，与相关社区和数据类型保持一致。该社区应对任何对创建或使用数据感兴趣的个人或组织开放。该组织应该由网络进行连接。

##### 9.2.2 数据库的维护

存储库的维护员应有一套明确定义的维护程序。包括维护网络，以及确保存储库满足其领域和使用的要求。

应有：

- a) 一个明确定义的过程，用于向存储库添加和删除数据；

- b) 一个明确定义的过程和计划，用于存储库的备份。

### 9.2.3 数据库结构

维护员应根据社区的需求定义存储库的模式。

存储库的数据模型应对整个存储库中的所有数据都适用，并与其他存储库可互相连接操作，例如基于现有的 F-A-I-R 模式基础设施（如适用）。

具体的模式由维护员和相关的数据社区决定。

### 9.2.4 与其他存储库的联系

维护员应尽可能与其他相关存储库进行链接。

### 9.2.5 授权和归属

维护员应发布：

- a) 存储库的授权模型；
- b) 存储库的归属模型。

### 9.2.6 稳定的 URI 和版本信息

维护员应提供一种机制：

- a) 用于创建和共享存储库中数据稳定的 URI；
- b) 用于对单个数据元素和记录以及存储库的整个版本进行版本控制。

### 9.2.7 数据可见性

维护员应使存储库中的数据对用户、常见的网络搜索和索引引擎可见和可访问。

维护员应使存储库中的数据可通过人工和/或程序访问，或两者都可行（例如，通过提供网络服务或数据下载）。可以应用有限的访问权限。

### 9.2.8 社区

维护员应将受存储库影响最大的社区纳入到创建、扩展和维护存储库的过程中。

### 9.2.9 语言

存储库中使用的人类语言应由维护组织自行决定。

维护组织应尽量使存储库在需要的情况下具备多语言版本。对于国际范围的使用情况，应提供控制词汇表和本体的英文版本。



## 附录 A

### (资料性)

### 生命科学数据常见格式示例

#### A.1 概述

附录A所列出的不同数据类型的格式并无排他性，使用类似的标准格式也可行。若依赖附录A中未列出的其他格式，只要满足所有其他规定和相关前提条件，即所应用的格式符合本文件的要求和建议，则仍符合本文件规定。对于需要快速适应相关数据格式的快速技术发展领域中的数据和模型格式，尤其适用。

可以在全球网络上找到有关生命科学数据格式的标准和词汇表的在线资源。其中一个广泛使用的示例是公开可用的FAIRsharing门户网站<sup>1)</sup>，它是一个经过审编、信息性和教育性的资源，涉及数据库和数据政策的数据和元数据标准。本文件中引用的推荐格式可以在“ISO 20691 FAIRsharing Collection”中找到，这是一个不断审编和更新的列表。除此之外，FAIRsharing和其他在线资源中还提供了比本附录所描述更多的格式和元数据格式。

本文件的主要范围并未包括医学影像、医疗数据记录、电子病历和其他与个人健康相关的数据格式，这些内容均在其他标准中作出了规定。然而需要强调的是，本文件中详述的通用数据类型的格式化和文档化建议，例如基因组数据方面的建议，不仅适用于医学领域，同时也适用于生命科学的其他领域，并且这些建议同样适用于健康数据领域的相关实践。尽管本文件可能并未涵盖医学数据格式化和文档化的所有方面，但其仍然可以为确保医学领域的互操作性以及提高数据质量提供极具价值的指导。

#### A.2 OMICS (组学)、生物化学和分子生物学方法的数据格式

##### A.2.1 蛋白质和核酸序列格式

###### A.2.1.1 概述

DNA和蛋白质序列使用了多种广泛认可的格式。所有这些序列数据格式均采用由国际生物化学与分子生物学联合会（IUBMB）以及国际纯粹与应用化学联合会（IUPAC）开发的命名法。

比较蛋白质的氨基酸序列可以确定它们之间的关系是否仅仅是偶然的。最初是针对共享密切相关功能的蛋白质进行视觉分析的，但是，这种分析缺少功能关系，无法解决直觉上的合理化问题。目前，视觉分析已经被提供简单对齐模型算法的统计方法所取代。对核苷酸序列进行比对、分析和特征提取等处理，也为转录、翻译和密码子选择提供了建模基础。

###### A.2.1.2 FASTA 序列格式

最常用于序列数据的格式是FASTA格式。FASTA格式是一种基于文本的格式，用于表示核苷酸序列或肽序列，其中核苷酸或氨基酸使用单字母代码表示。该格式还可以在序列前附带序列名称和注释。尽管该格式较为简单，但并不具备扩展注释信息的功能。使用者可以轻松获得能够实现从无格式文件和其他格式转换的软件，此类高效的转换能够避免涉及本体论的陷阱。

###### A.2.1.3 FASTQ 序列和序列质量格式

1) 本文档提供此信息是为了方便用户，这并不构成ISO对该产品的认可。其他网络资源也可用于总结适合的格式和词汇表。

FASTQ是由桑格研究所开发的一种文本文件格式，它能够在单个文件中同时包含测序序列和每个碱基的质量值，从而实现了测序数据的共享。不同测序平台所采用的FASTQ格式具有标量差异，但它们之间是可以相互转换的。

FAST5格式基于分层数据格式HDF5格式，可用于存储大量复杂数据。

例如，FAST5格式是Oxford Nanopore测序仪<sup>2)</sup>（如MinION）在标准测序输出。

与fasta和fastq文件不同，FAST5文件是二进制文件，不能使用普通文本编辑器打开。

存储在nanopore FAST5文件中的数据可以包含一段（经过碱基识别处理后）的fastq格式序列数据和纳米孔的原始信号信息，以及一些日志文件和其他信息。

#### A. 2. 1. 4 序列读取格式 SRF

国际核苷酸序列数据库共享联盟（INSDC）是由日本DNA数据库（DDBJ）、欧洲生物信息研究所（EMBL-EBI）的欧洲核苷酸数据归档库（ENA）和美国国家生物技术信息中心（NCBI）的测序数据归档库（SRA）组成，其中SRA是美国国立卫生研究院（NIH）的高通量测序数据主要归档库。NCBI是高通量测序数据的主要存档库。INSDC设计了DDBJ/ENA/GenBank特征表，该表综合纳入了EMBL、GenBank和DDBJ这三个数据库的序列格式。向这三个数据库提交的数据将在它们之间进行共享。

相比毛细管凝胶电泳产生的数据文件（通常只包括单一的图线和一些元数据），大规模并行测序产生的数据文件中没有包括元数据，但可以进行额外分析且更加实用。SRA数据库已经将元数据剥离开，并能够支持多种测序平台的数据格式。一些专有数据格式的转换效率仍在不断提高，SRF已开始解决这种转换问题。SRF是一种用于DNA序列数据的通用格式，主要目的是使其能够作为存储任何DNA测序技术生成的数据的单一格式。因此，该格式具有足够的灵活性，可以以最小的实现成本存储当前和未来DNA测序技术的数据。

#### A. 2. 1. 5 序列注释格式

- a) （基因组）浏览器可延展数据（BED）格式是一种灵活的格式，用于定义 UCSC 基因组浏览器注释轨迹中显示的每行数据。BED 格式文件中的每行数据包括三个必填字段和九个可选字段。在注释轨迹的任何单个数据集中，每行字段的数量必须保持一致。BigBED 轨迹格式存储可以是简单的，也可以链接到一系列外显子的轨迹注释。BigBED 文件是 BED 文件的集合。该格式由全球基因组和健康联盟（GA4GH）维护。
- b) （基因组）浏览器图形绘制数据（WIG）格式是一种面向行的格式，用于在 UCSC 基因组浏览器中绘制图形。WIG 已经被 bigwig 取代。
- c) 通用特征格式（GFF3）是通用模式生物数据库最新的可用版本，解决桑格 Sanger 研究所以前设计的版本缺陷。它有九个制表符分隔的字段。它采用九个制表符分隔的字段来表示各种类型的数据，包括：典型基因、非编码转录本、父代（部分）关系、比对、本体关联信息和数据库间的交叉引用、单外显子基因、多基因转录本、包含内嵌子的基因、跨剪接转录本、程序化移码和操纵子。
- d) 变异位点格式（VCF）及其二进制版本二进制变异位点格式（BCF）是文本文件格式，通常以压缩形式存储序列变异数据文件。这些文件格式可以包含元信息行、标题行和数据行。该格式由全球基因组和健康联盟（GA4GH）维护。变异位点表示相对于某个参考位点的 DNA 序列变化。通常情况下，VCF 文件中的一行对应于一个变异位点，这些变异位点可能表示单核苷酸多态性（SNP）或插入等不同类型的变异。

2) 这些信息是为了方便本文档的用户而提供的，并不构成ISO对该产品的认可。它仅作为FAST5格式来源的具体示例。

- e) 基因转移格式 (GTF) 借鉴了 GFF 的结构, 但具有一些额外的信息结构, 需要单独定义和格式名称。结构与 GFF 相同, 因此字段包括:  
 <seqname 序列名称> <source 数据源> <feature 特征类型> <start 特征的起始位置> <end 特征的终止位置> <score 分数> <strand 链向> <frame 相对于整个编码区域的第一个碱基的偏移量> [attributes 其他属性] [comments 备注]
- f) 基因组变异格式 (GVF) 是对通用特征格式第三版 (GFF3) 的扩展, 该格式采用了简单的制表符分隔, 专门用于描述 DNA 变异数据文件, 利用序列本体对基因组变异数据进行详细描述。
- g) 合成生物学开放语言 (SBOL) 是一种用于表示基因回路设计等序列的 RDF 格式。它具有丰富的信息表达能力, 可以表达序列特征注释和部件/子部件关系。此外, SBOL 还可以用于表示不完整/部分序列和基因设计中部件的相对顺序。

#### A. 2.1.6 序列数据压缩格式

CRAM 是一种压缩的列式文件格式, 用于存储比对到参考序列的生物序列, 最初由 Markus Hsi-Yang Fritz 等人设计。CRAM 作为基于参考序列的压缩格式, 旨在成为序列比对格式 (SAM) (参见 A. 2.2.3) 和二进制序列比对格式 (BAM) (参见 A. 2.2.4) 的有效替代。它可以选择使用基因组参考序列来描述序列片段与参考序列之间比对后的差异信息, 从而降低存储成本。此外, SAM 格式中的每一列都被单独分成块, 从而提高压缩率。CRAM 文件通常比 BAM 文件小 30% 到 60% 不等, 具体取决于其中包含的数据。CRAM 格式规范由全球基因组和健康联盟 (GA4GH) 维护。

ISO/IEC 23092 系列 (MPEG-G) 是国际标准化组织 (ISO) 和国际电工委员会 (IEC) 制定的一系列用于基因组信息表示的标准。该系列标准旨在为由高通量测序仪生成的数据信息及其后续处理和分析的不同可能实现提供可互操作的数据存储、访问和保护解决方案。该系列标准利用了数字媒体领域先前验证过的技术和数据表示架构。它们允许压缩和传输基因组测序数据, 即使在复杂的情况下, 例如在需要访问大量可能分布的数据时, 或者在需要对部分数据进行加密以保护隐私时。该系列标准由不同的部分组成, 每个部分都解决一个特定的方面, 例如压缩、元数据关联、应用程序编程接口 (API) 和用于数据解码的参考软件。

- a) ISO/IEC 23092-1 规定了基因组信息的传输 (例如流式传输) 和存储的数据格式, 包括转换过程。
- b) ISO/IEC 23092-2 规定了几种类型基因组信息的表示规范, 例如测序数据的 MPEG-G 无损压缩和关联质量分数的有损压缩的语法和方法。该标准仅规定了解码过程和解码器输出格式, 而编码过程则留给算法和实现特定的创新。
- c) ISO/IEC 23092-3 规定了元数据的存储和解释, 以及为 ISO/IEC 23092-1 中规定的不同封装级别提供机密性、完整性和隐私规则的保护元素, 并定义了用于访问符合 ISO/IEC 23092-1 和 ISO/IEC 23092-2 编码的基因组信息的 API。此外, 该标准还包括了有关如何将辅助字段与编码读取相关联的规范以及与现有 SAM 内容 (见 A. 2.2.3) 向后兼容的机制, 以及导出到此格式的机制。
- d) ISO/IEC 23092-4 规定了基因组信息表示参考软件, 称为“基因组模型”。该标准提供了解码软件以评估其是否符合 ISO/IEC 23092-1、ISO/IEC 23092-2 和 ISO/IEC 23092-6 的要求。
- e) ISO/IEC 23092-5 规定了一组测试程序, 旨在验证比特流和解码器是否满足 ISO/IEC 23092-1 和 ISO/IEC 23092-2 中规定的要求。该标准识别这些要求, 将其与测试功能相关联, 并定义了如何测试对其的符合性。根据这些功能实现的测试比特流以电子形式提供。
- f) ISO/IEC 23092-6 规定了基因组信息注释的规范表示和编码, 例如具有基因型信息的变异、功能性注释、轨迹、表达矩阵和接触矩阵。

#### A. 2. 1. 7 加密基因组数据的格式

Crypt4GH是一种用于以加密和认证的方式存储基因组数据的文件容器格式，例如BAM或CRAM。该方法使用双重信封加密：数据本身被加密，解锁数据的机制也被加密。收件人必须有自己的私钥来验证其身份，并需要特定密钥来访问传输文件中的数据。Crypt4GH格式规范由全球基因组和健康联盟（GA4GH）维护。

#### A. 2. 2 序列比对格式

##### A. 2. 2. 1 一般情况

序列比对通常是进行功能比较的基础。

##### A. 2. 2. 2 CLUSTAL-W

CLUSTAL-W对齐格式是一种简单的文本基础格式，通常使用\*.aln文件扩展名，用于将DNA或蛋白质序列输入和输出到Clustal套件的多重比对程序中。Clustal W已经得到了很好的开发和许多应用程序的支持。

##### A. 2. 2. 3 序列比对/映射（SAM）

SAM是一个由制表符分隔的文本格式，包括可选的头部部分和比对部分。该格式专门设计用于处理大量序列。如果存在头部部分，则必须在比对部分之前。头部行以“@”开始，而比对行则不以“@”开始。每行比对具有11个必填字段，用于提供基本的比对信息，例如映射位置，以及可变数量的可选字段，用于存储灵活或特定于比对的信息。

##### A. 2. 2. 4 二进制比对映射（BAM）格式

BAM是序列比对/映射（SAM）格式的压缩二进制版本，是核苷酸序列比对的紧凑且可索引表示形式。许多新一代测序和分析工具都使用SAM/BAM格式。对于自定义轨迹展示，经过索引的BAM相比于PSL和其他人类可读的比对格式的主要优点在于，只有用于展示特定区域所需的文件部分会被传输至UCSC。这使得展示来自巨大文件的比对成为可能，因为如果试图将整个文件上传至UCSC，可能会导致与UCSC的连接超时。BAM文件及其相关的索引文件保存在您可从网络访问的服务器（http、https或ftp）上，而非UCSC服务器。UCSC会临时缓存所访问的文件部分以加快交互式展示速度。

##### A. 2. 2. 5 基于快速傅里叶变换的多重比对（MAFFT）

MAFFT是一种高速多序列比对程序，该程序基于快速傅里叶变换（FFT）方法，根据氨基酸的物理特性优化蛋白质序列比对。该程序使用渐进和迭代两种比对方法。MAFFT适用于难比对的序列，例如包含大空缺的序列（例如包含可变环区的rRNA序列）。同时，FASTA和Pearson格式也适用于MAFFT。

##### A. 2. 2. 6 斯德哥尔摩多重比对格式

“斯德哥尔摩”格式是一种用于多序列比对的数据文件格式，用于存储比对特征。在这种格式中，每个特征前面有一个“magic”标签，总共有四种类型。斯德哥尔摩格式可以被HMMER、Pfam和Belvu等程序使用。

##### A. 2. 2. 7 其他序列比对格式

其他格式，例如FASTA、phyllip和多序列格式（MSF）等。可以使用接口进行这些格式之间的相互转换。

### A.2.3 RNA序列、结构和连接格式

几种用于存储RNA结构数据的格式，例如，BIOpolymer Markup Language (BIOML)、Bioinformatic Sequence Markup Language (BSMLTM)、Genome Annotation Markup Elements (GAME) 和 CORBA Bio effort。这些格式没有标准的语法。RNAML (RNA Markup Language) 由RNA生物信息学领域具有相当广泛代表性的研究人员联合开发，被广泛用于RNA信息文件。其中“.ct文件”包含了核酸序列和碱基配对信息，这些信息可以用来计算出结构图谱。

目前存在几种不同的“.ct”格式变体，用于不同的应用程序。

——点括号文件格式 (DBN)：RNA 二级结构通常使用点-括号符号 (DBN) 格式来描述。有效的 DBN 格式结构是由点“.”、开括号“(”和闭括号“)”构成的带括号的单词。点表示未配对的核苷酸，匹配的括号表示配对的核苷酸。由于相互作用的核苷酸数始终是偶数（每个核苷酸都是以配对的形式存在），因此括号必须成对出现。如果一个结构中至少包含两个茎环结构，其中一个茎的一半插入到另一个茎的两半之间，这样的结构被称为“假结”。假结首次在 1982 年的芜菁黄化花叶病毒中得到确认。假结折叠成纽结状的三维构象，但它们并不是数学意义上的真正拓扑纽结。假结使用替代的[...]或{...} 括号对进行标记。

——BPSEQ：BPSEQ 是一种文件格式，文件名以“.bpseq”结尾。BPSEQ 格式采用简单的文本格式，其中每行表示一个碱基分子，包含该碱基的位置编号（最左边的位置为 1）、碱基的名称（A、C、G、U 或其他字母），以及与其配对的碱基的位置编号，如果该碱基未配对，则编号为 0。更多信息请参见比较 RNA 网站 (Comparative RNA Web Site)。

对于包含多个分子的复合物，这些分子按顺序列出，每个连续分子的碱基对编号都紧随前一个分子。

——CT：第一行包含序列长度 L。后续共有 L 行，每行代表一个核苷酸。第 i 行首先以 i 开始，接着是代表第 i 个核苷酸的字母，然后是 5'-连接的碱基索引 (i-1)，接着是 3'-连接的碱基索引 (i+1)，接着是配对的碱基索引（如果未配对则为 0），最后是原始序列中的碱基索引。

例如，上述以 bpseq 格式表示的结构以“.ct”格式表示如下：

```
8
1 G 0 2 8 1
2 G 1 3 7 2
3 C 2 4 0 3
4 A 3 5 0 4
5 U 4 6 0 5
6 U 5 7 0 6
7 C 6 8 2 7
8 C 7 0 1 8
```

CT 格式适用于两个或更多 RNA 分子的复合物，因该格式可明确表达边界信息。如果第 i 行对应于一个分子中的第一个核苷酸，则第三列是 0。如果第 i 行对应于一个分子中的最后一个核苷酸，则第四列是 0。

这是一个短的双链茎，即由两个分子组成的茎：

```
4
1 G 0 2 4 1
2 G 1 0 3 2
3 C 0 4 2 3
4 C 3 0 1 4
```

——RNAML：该格式的语法能够存储和交换 RNA 序列、二级结构和三级结构的信息。同时，它还允许描述更高层次的数据信息，这些信息包括但不限于碱基对、碱基三元组以及假结。面向类的方法使我们能够表示一组 RNA 分子共有的数据，例如序列比对和共有的二级结构。

## A. 2. 4 质谱数据格式

### A. 2. 4. 1 质谱仪输出文件格式（mzML）

mzML是LC-MS蛋白质组学数据的通用格式，由HUPO-PSI工作组针对质谱标准制定的mzXML和mzData逐步演进而来。该格式已经得到了很好的发展和“防过时”验证。ProteoWizard软件项目拥有许多工具，可以对不同平台的XML进行转换，包括根据相关许可条款而变化的厂商专有软件。稳定的版本mzML 1.1自2009年以来就已经发布。

### A. 2. 4. 2 基于质谱的蛋白质组学定量研究（mzQuantML）

mzQuantML标准格式旨在存储通过质谱对分子（主要是肽和蛋白质）进行定量的工作流程的系统描述。该格式最初以AnalysisXML的名称为蛋白质组学背景下质谱的几种计算分析类型开发，但随后决定分为两种格式：用于肽和蛋白质鉴定的mzIdentML和用于分子定量的mzQuantML。MzIdentML是蛋白质组学标准化倡议（PSI）蛋白质组学信息学工作组所指定的标准之一，并不包含质谱数据，需要额外提供。

TraML是用于靶向质谱方法定义的丰富XML格式的标准。TraML基于与mzML和mzIdentML相同的设计理念。像先前为不同数据类型开发的这些格式一样，TraML基于XML，可以使用许多行业标准工具解析和验证其结构正确性。

proBAM和proBed格式旨在存储以基因组为中心的蛋白质组学数据的表示。

## A. 2. 5 蛋白质组学和代谢组学数据交换格式（mzTab）

mzTab的开发旨在弥合XML格式所需的建模完整蛋白质组学数据的高层次细节信息与蛋白质组学和代谢组学数据的必要信息之间的差距。这个来自HUPO-PSI制定的此格式非常适合将基于质谱的蛋白质组学和代谢组学结果提供给质谱（MS）领域之外更广泛的生物界使用。mzTab解决了许多结构和功能各不相同的蛋白质中存在特定肽序列的情况。

## A. 2. 6 核磁共振光谱格式（NMR）

NMR-star是核磁共振光谱NMR数据的首选格式。它是自我定义文本存档和检索（STAR）文件的扩展，或简称为STAR文件。NMR-star是一种用于存储结构化数据的基于文本的文件格式。其他格式包括PIPP和XEasy。

## A. 2. 7 酶动力学数据格式：EnzymeML

EnzymeML是一种免费且开放的基于XML的标准交换格式，用于存储有关酶催化反应的数据。EnzymeML的目的是在仪器、软件工具和数据库之间存储和交换酶动力学数据。即使科学家们使用不同的仪器、电子实验室笔记本或数据库，EnzymeML也能让他们分享实验方案和结果。EnzymeML与系统生物学标记语言（SBML）兼容。它由国际社区不断发展和扩展。

## A. 2. 8 凝胶电泳数据格式：凝胶标记语言（GelML）

凝胶电泳通常依赖于许多物理参数，这些参数以化学、构象和电磁的方式在一个或两个维度中有所差异。

在凝胶电泳中，元数据具有非常重要的意义。由HUP0-PSI开发的Ge1ML是用于表示在蛋白质组学研究中凝胶电泳实验数据的交换格式。按照蛋白质组学实验的最低信息（MIAPE）要求，Ge1ML保留了功能基因组学实验（FuGE）对象模型中的几个结构。

### A. 2. 9 实时PCR数据格式

MIQE（定量实时PCR 实验发布的最低限度信息）是一套指南，用于描述评估定量实时PCR实验（qPCR）所需的最基本信息。目前，该格式也适用于数字PCR（dPCR）。

### A. 2. 10 基因组测序数据格式

MIxS（关于任何（X）序列的最小信息）指南由基因组标准联盟（GSC）开发，旨在提高基因组序列数据的可发现性，从而实现数据整合、发现和比较。

MIxS提供了描述基因组、宏基因组和基因标记序列的核心标准。它是MIGS（关于基因组序列的最小信息）和MIMS（关于（宏）基因组序列的最小信息）的扩展。目前，它由三个独立的检查表组成：用于基因组的MIGS、用于宏基因组的MIMS和用于标记基因的MIMARKS。

### A. 2. 11 生物大分子结构数据

生物大分子晶体学信息文件格式（PDBx/mmCIF）是用于存档生物大分子晶体学实验及其结果的数据字典。它被用作官方发布格式，用于格式化和交换核酸和蛋白质等生物大分子的结构数据。这些结构最初以自蛋白质数据库（PDB）的三种晶体学数据格式中的一种进行格式化。通常使用标准制表符分隔格式提供原子坐标和文献条目格式（PDB，1992；PDB，1996）。PDB交换数据字典（PDBML）提供XML模式。

PSI扩展FASTA格式（PEFF）是一种用于蛋白质和核苷酸序列数据库的统一格式，可供序列搜索引擎和其他相关工具（光谱库搜索工具、序列比对软件、数据存储库等）使用。该格式可以在不同软件平台上一致地提取、显示和处理蛋白质/核苷酸序列数据库条目标识符、描述、分类等信息，并允许表示结构注释，如翻译后修饰、突变和其他。该格式采用扩展FASTA格式的纯文本文件形式，其中包含描述序列条目的元数据头，以描述从数据库中获取的相关信息（即名称、版本等）。

### A. 2. 12 小分子（化学实体）结构

IUPAC国际化学标识符（InChI）是用于化学物质的文本标识符，旨在提供一种标准的编码方式来描述分子信息，并促进在数据库和网上搜索这些信息。该格式最初由国际纯粹与应用化学联合会（IUPAC）和美国国家标准与技术研究院（NIST）在2000年至2005年开发，格式和算法是非专有的。InChI标识符根据信息层来描述化学物质：原子及其键连接、互变异构体信息、同位素信息、立体化学和电子电荷信息。并非所有层都必须提供；例如，如果同分异构体信息与特定应用无关，则可以省略同分异构体层。InChIKey是完整InChI的哈希版本，具有固定长度（27个字符）的InChI压缩数字表示形式，不可被人类直接理解。InChIKey规范于2007年9月发布，以方便对化学化合物的网络搜索。与InChI不同，InChIKey不是唯一的，但现在还没有已知的InChIKey碰撞（尚未发现具有相同InChIKey的两个不同结构）。InChI格式在文献和生物信息学应用中广泛使用。

简化分子线性输入规范（SMILES）是SMILES语言的开放框架版本，是一种用于指定化学结构的排版行符号。它在Blue Obelisk项目的旗帜下托管，目的是征求整个计算化学界的贡献和意见。OpenSMILES是社区赞助的SMILES开放标准版本。SMILES在文献和生物信息学应用中也广泛使用。

MDL Molfile是一种文件格式，用于保存有关分子的原子、键、连接性和坐标的信息。每个molfile文件描述一个单独的分子结构，可以包含不连通的片段。V3000 molfile和V3000 rxnfile格式是规范的最新版本。V3000是以不同格式的V2000的超集。连接表（Ctab）包含描述一组原子的结构关系和属性的信息。尽管通常被称为molfile，但随着最初创建并拥有该文件格式规范的公司的更迭，规范的所有权

也发生了变化。MDL信息系统公司（MDL Information Systems）最初开发了该格式，目前由达索系统（Dassault Systemes）拥有。请注意，无法找到Molfile的官方公开主页，因此只提供了其他的替代信息。

结构数据格式（SDF）是一种化学文件格式，用于表示多个化学结构记录和相关的数据字段。SDF是由Molecular Design Limited（MDL）开发和发布的，成为最广泛使用的化学品信息导入和导出标准。

#### A. 2.13 用于基于测序的功能基因组学的微阵列数据和格式

功能基因组数据学会（FGED Society）最初推出了一种名为MAGE-ML的微阵列数据标记语言格式。后来发现，这款产品对于没有生物信息学知识基础的研究者来说过于复杂。因此，开发出了MAGE-Tab。它是一种基于电子表格的微阵列基因表达数据格式。MicroArray基因表达制表符（MAGE-TAB）格式是独立的，不需要理解MAGE-ML或XML。MAGE-TAB可以用于按照MIAME指南存储微阵列数据，或者按照MINSEQ指南存储测序实验数据。

MIAME（关于微阵列实验的最小信息集）旨在明确规定对微阵列实验进行准确描述所需的所有必要信息，并让重复实验结果成为可能。MIAME定义了这些信息的内容，但没有规定格式。MIAME/Plant（关于植物微阵列实验的最小信息指南）是MIAME指南的一个扩展，描述了在涉及植物的微阵列实验时应记录哪些生物学细节信息。需要提供关于生物方面的详细信息，如生长条件、收获时间或收获器官等。

MINSEQE（高通量测序实验的最小信息集）描述了高通量核苷酸测序实验所需的最小信息，以使人们能够明确解释并方便重复实验结果。类似于用于微阵列实验的MIAME指南，遵守MINSEQE指南将改善不同模式的多个实验的整合，从而最大限度地提高高通量测序研究的价值。

#### A. 2.14 糖组学数据格式——糖组学实验所需的最小信息——基于质谱的糖组学数据分析（MIRAGE MS）

MIRAGE（糖组学实验所需的最小信息集）的创建旨在改善科学文献中糖组学数据的质量。为了更好地理解碳水化合物的生物化学结构-功能关系，研究人员需要提供详细的测定条件和实验结果描述。然而，目前在文献中对这些数据的报告不够充分。至关重要的是对样本制备工作流程进行基本描述。与蛋白质组学不同，不同类型的糖基化基团物需要采用部分不同的释放方法，而这些方法可能直接影响以下条件/参数：已释放的糖或仍附着在蛋白质/脂质上的糖；糖的类型（N-糖基化、O-糖基化、蛋白多糖片段）；以及在进行质谱分析（MS）之前进行的样本前处理（非、还原、甲基化、内/外糖酶消化、荧光标记、在线/离线液相色谱分离）。

#### A. 2.15 流式细胞术实验数据的格式——关于流式细胞术实验的最小信息集（MIFlowCyt）

流式细胞术实验的最小信息集（MIFlowCyt）建立了有关流式细胞术实验概述、样本、仪器和数据记录的记录和报告信息的标准。它通过指定数据内容的要求并提供捕获信息的结构化框架，促进了对流式细胞术实验的临床、生物和技术问题进行一致注释。

#### A. 2.16 描述合成生物学元件、部件和系统的数据格式

##### A. 2.16.1 合成生物学开放语言（SBOL）

SBOL是一种开放标准，用于表示基于计算机的生物设计，以及通过设计-构建-测试-学习工作流程来实现这些设计。SBOL数据提供了以电子格式表示这些信息（SBOL），并使用图形符号来形象地描述遗传设计（SBOL Visual）。

##### A. 2.16.2 合成生物学开放语言视觉版（SBOL Visual）



合成生物学开放语言视觉版（SBOL Visual）是一种开源的图形符号，使用示意图“符号”来表示遗传元件、部件、模块和系统。

#### A. 2. 17 生命科学、环境和生物医学实验的元数据格式

##### A. 2. 17. 1 基于调查/研究/分析模型的制表分隔格式（ISA-TAB）

ISA-Tab描述了使用ISA-Tab格式指定的ISA（调查/研究/分析）抽象模型的参考实现。ISA-Tab文件是制表符分隔值（tsv）文件，具有特定的列结构。ISA模型由三个核心实体组成，用于捕获实验元数据：调查、研究和分析。该模型的可扩展、分层结构使其能够表示使用一种或多种技术的研究，重点在于描述其实验元数据（即样本特征、技术和测量类型、样本到数据关系）。

##### A. 2. 17. 2 基于调查/研究/分析模型的 JSON 格式（ISA-JSON）

ISA-JSON描述了一种使用JSON格式的参考实现，用于规范ISA抽象模型。JSON是一种用于序列化结构化数据的文本格式。ISA模型由三个核心实体组成，用于捕获实验元数据：调查、研究和分析。该模型的可扩展、分层结构使其能够表示使用一种或多种技术的研究，重点在于描述其实验元数据（即样本特征、技术和测量类型、样本到数据关系）。

##### A. 2. 17. 3 快捷式健康照护互操作性资源（FHIR）

FHIR旨在支持各种场景下的医疗保健信息交换。该规范基于并适应各种实践，以使广泛的团队和组织能够提供综合医疗保健。FHIR的预期范围很广，涵盖人类和兽医、临床护理、公共卫生、临床试验、行政管理和财务方面。该标准由HL7开发，并且旨在全球范围内广泛使用，适用于各种体系结构和场景。

#### A. 3 生物成像数据的格式

##### A. 3. 1 影像数据资源格式（IDR）

IDR是一个用于发布、挖掘和集成大规模生物成像数据的原型平台，遵循欧洲生物医学影像基础设施（Euro-BioImaging）/ELIXIR成像战略，使用由开放显微镜环境项目（Open Microscopy Environment）构建的OMERO和Bio-Formats开源软件。该平台部署在EMBL-EBI的Embassy资源上运行的OpenStack，包括与遗传学、RNAi、化学、地理位置高内涵筛选、超分辨率显微镜和数字病理学等独立研究中的图像数据。

##### A. 3. 2 开放式显微镜环境可扩展标记语言格式（OME-XML）

开放式显微镜环境项目（OME）为了生物显微镜数据的存储和处理开发了相关的开源软件和数据格式标准。该项目是由大学、研究机构、工业界和软件开发团体联合推进。OME-XML的目的是提供一种丰富的、可扩展的方法来保存与显微镜实验及其中获取的图像有关的信息。

##### A. 3. 3 开放显微镜环境本体论模型（OME-OWL）

开放式显微镜环境本体论模型（OME-OWL）是一个通过OME数据模型的转换而开发的轻型显微镜成像本体论模型。该本体论模型的目的是支持多模态成像技术，并集成生命科学领域的元数据，以便进行全面的图像分析。从OME数据模型中提取的核心概念包括项目、实验、仪器、图像、屏幕、平板和感兴趣区域。该本体论模型已经扩展到包括电子显微镜、X射线计算机断层扫描（CT）和磁共振成像（MRI）。

#### A. 4 应用于生物系统计算机模型的数据格式

##### A. 4. 1 CellML

CellML是一种机器可读的、基于XML的模型描述和交换格式，用于基于计算机的数学模型。它是一种描述语言，用于定义细胞和亚细胞过程的模型。它定义了将数学关系分组模块中的轻量级XML结构。数学中使用的变量在每个模块中定义，并且可以指定不同模块之间变量之间的连接。CellML支持基于组件的建模，允许模型导入其他模型或模型的部分组件，因此强烈鼓励模型的重复使用，并促进模块化建模方法。CellML模型通常由可以包含描述每个组件行为的变量和数学的组件组成。该格式提供了将组件重用和组合成层次结构的方法。所有实体（元素）都带有标识符，数学定义使用MathML进行编码。数学模型被认为是主要数据，生物学相关的上下文信息是通过使用RDF对变量和方程注释元数据来提供的。

#### A. 4.2 系统生物学标记语言（SBML）

SBML一种基于XML的机器可读模型描述和交换格式，用于生物过程的计算模型。其优点在于表示生化过程规模的现象，但并不仅限于此。SBML的发展分阶段进行，每个“层次”都试图在一定的复杂性级别上实现一致的语言。由于SBML Level 3的格式是模块化的，因此核心可以独立使用，而包则是向核心额外添加的功能“层”。就其本身而言，SBML核心适合表示诸如经典代谢模型和细胞信号模型等事物，涉及良好混合的物质和它们所在的空间均匀区隔。扩展核心并可选择使用的SBML包增加了额外的模型特性，如可视化、基于约束的模型、层次模型组合或元素分组。SBML模型被分解为显式标记的组成元素。一个有效的模型可以由各种用户定义的元素组成，例如，过程中涉及的物质、产品和修饰剂，或隔间及其位置。

使用RDF对模型、其参数、变量、方程、实体、组件和其他内容进行注释，可以提供模型的生物和结构上下文信息，以及模型的内容和环境信息。

#### A. 4.3 神经科学可扩展标记语言（NeuroML）

NeuroML是一种机器可读、基于XML的模型描述和交换格式，专门用于神经科学计算机模型。它的创建旨在促进神经科学领域的数据存档、数据和模型交换、数据库创建和模型发布，重点关注基于真实神经元的生物物理和解剖学特性的模型。

#### A. 4.4 药物计量学标记语言（PharmML）

PharmML是一种用于药效学中的非线性混合效应模型的交换格式，提供一种编码模型、试验设计和建模步骤的方法。PharmML允许在群体药代动力学和药效学中使用的不同软件工具之间平稳交换计算机模型。

#### A. 4.5 人类生理组场标记语言（FieldML）

FieldML是一种机器可读、基于XML的模型描述和交换格式，专门用于表示分层模型的广义数学场。FieldML能用于3D动态几何等方面的信息，来描述细胞、组织和器官等计算模型。

#### A. 4.6 生物路径交换（BioPAX）

BioPAX是一种机器可读的标准格式，其主要目的是实现生物途径数据的集成、交换、可视化和分析。

#### A. 4.7 系统生物学图形符号（SBGN）

系统生物学图形符号（SBGN）是一项用于规范生物过程图解中使用的图形符号的计划。SBGN的目标是开发高质量、标准图形语言，用于表示生物过程和相互作用。使用标准的可视化符号对于确保示意图和代谢通路图是明确和一致的，具有极其重要的意义。SBGN包括三种语言，这些语言从不同的角度和细节水平上观察生物网络。

- SBGN 过程描述语言（PD）：SBGN 过程描述（PD）语言显示网络中生化相互作用的时序过程。它可用于展示网络中所有生化实体的分子相互作用，同一实体可在同一张图中多次出现。
- SBGN 实体关系语言（ER）：SBGN 实体关系（ER）语言允许查看给定实体参与的所有关系，不考虑时间因素。关系可视为描述实体节点对其他关系影响的规则。
- SBGN 活动流程语言（AF）：SBGN 活动流程（AF）语言描述网络中生化实体之间的信息流。它省略了有关实体状态转换的信息，特别适合表示遗传或环境等扰动的影响。

## A.5 应用于生命科学模型模拟及其结果的数据格式

### A.5.1 模拟实验描述标记语言（SED-ML）

模拟实验描述标记语言（SED-ML）是一种机器可读的，基于XML的数据交换格式，用于编码计算模型仿真设置信息，以确保仿真实验数据的可交换性和可重复性。它遵循MIASE准则中定义的要求（详见条款B.2）。

SED-ML允许配置和重新运行精确的仿真设置。此数据格式按照不同的级别和版本进行演化，同时不依赖于任何特定的仿真软件或建模格式。

### A.5.2 开放模型交换格式（OMEX）

开放模型交换格式（OMEX）支持交换生命科学中建模和仿真实验所需的所有信息的数据格式。OMEX文件是一个ZIP容器，包括清单文件、可选的元数据文件以及描述模型的多个文件。清单文件是一个XML文件，列出了存档中包含的所有文件及其类型。元数据文件提供了有关存档及其内容的额外信息。虽然可以使用任何格式，但使用XML序列化的RDF格式是最佳选择。

### A.5.3 数值标记语言（NuML）

数值标记语言（NuML）是一种机器可读的，基于XML的数据交换格式，用于描述和交换模型和仿真描述中的多维数值数组。NuML最初作为系统生物学结果标记语言（SBRML）的一部分而被开发。

## A.6 用于数据和模型质量测量的描述符

质量控制标记语言（qcML）是一种机器可读的，基于XML的数据交换格式，用于质谱法中与质量相关的数据，遵循HUPO-PSI（蛋白质组学标准倡议）相关标准 mzML、mzIdentML、mzQuantML 和 TraML 的设计原则。这种数据格式，主要用于生命科学中高通量实验和过程中获得的质量控制（QC）数据。该项目目前的重点是基于质谱法的蛋白质组学，但该格式也适用于代谢组学和下一代测序。

附录 B  
(资料性)  
数据、模型和元数据的最低报告标准

B.1 概述

附件B提供了适用于生命科学领域的最小信息标准清单以及适用于特定领域的本体、分类和受控词汇表清单，以描述数据集及其包含的数据元素和上下文。附录B中列出的报告标准并非排他性的，也可以使用其他类似标准（如果适用）。

除所列标准外，只要满足本文件中的所有其他要求及相关前提条件，其他标准也可使用（见第4至第8条）。

有关生命科学数据格式的标准和词汇表的在线资源可在全球的网络上找到。其中一个广泛使用的示例是公开可用的FAIRsharing门户网站<sup>3)</sup>，它是一个经过审编、具有信息性和教育性的资源，涉及到数据库和数据政策的数据和元数据标准。附录B.2中列出的最低信息标准和附录B.3中列出的术语可以在“ISO 20691 FAIRsharing Collection”中找到，这是一个不断审编和更新的列表。在FAIRsharing和其他在线资源中，还可以找到比本附录描述的更多的格式和元数据格式。

附录B.3中的本体论列表是从NCBO BioPortal和EMBL-EBI Ontology Lookup Service检索得到的一组初始资源，根据它们在生命科学数据管理资源中（如广泛使用的SEEK平台及其FAIRDOMHub安装程序等）中的积极使用情况而被认为具有相关性。

B.2 最低报告标准

表 B.1 表最低信息标准清单

首字母缩写	名字	描述和主页
CIMR	代谢组学报告的核心信息	代谢组学实验的最低要求。
CONSORT	试验报告综合标准	平行组随机对照试验（RCT）的报告，使读者能够了解试验的设计、实施、分析和解释，并评估其结果的有效性。
MIABE	关于生物活性实体的最低限度信息	针对一种或一系列生物活性实体（如药物和农药）及其与一个或多个靶点分子的相互作用数据的发表的报告要求。
MIABIS	生物库数据共享的最低信息要求	生物库之间的合作以及促进生物样本和数据交换所必需的最小信息。本标准的目的是通过协调生物库和生物医学研究，促进生物资源的重复利用和相关数据的互用性。
MIACA	细胞分析的最低信息要求	信息指南和模块化的细胞测定对象模型（CA-OM），可涵盖各种可能的细胞分析，并为高效的数据交换提供基础。
MIAME	微阵列实验的最低信息要求	使实验结果的解释明确且可能重现该实验所需的信息。
MIAPAR	蛋白亲和试剂的最低信息要求	这是一个为希望明确描述蛋白质亲和试剂及其蛋白质靶标的实验人员提供的指南。它规定了描述亲和试剂（如抗体、适体、蛋白质支架等）的生产或使用所需的最小信息集合。
MIAPA	系统发育分析的最低信息要求	研究人员用于评估或重复使用已发表的系统发育分析的必要元数据清单。

3) 本信息仅为方便本文件的用户而提供，并不构成ISO对本产品的认可。其他网络资源也可用于总结合适的格式和词汇。

MIAPE	蛋白质组学实验的最低信息要求	公共归档库所需的关于整个蛋白质组学实验的最小信息集合。
MIAPE-MS	蛋白质组学实验的最低信息要求-质谱学	蛋白质组学实验中使用质谱仪所需的最小信息集合，这些信息应足以支持对数据的（重新）解释和（重新）评估，以及可能重现产生这些数据的实验操作。
MIAPE-MSI	蛋白质组学实验的最低信息要求-质谱信息学	为分析质谱实验产生的数据而使用蛋白质和肽鉴定和表征软件所需的最小信息集合，这些信息应足以支持对数据的有效解释和评估以及可能重现产生这些数据的实验操作。
MIAPE-GE	蛋白质组学实验的最低信息要求-凝胶电泳	在蛋白质组学实验中使用N维凝胶电泳的最小信息集合。
MIAPE-GI	蛋白质组学实验的最低信息要求-凝胶信息学	以符合“MIAPE原则”文件规定的方式，报告利用凝胶电泳图像进行的信息分析所需的最小信息集合。
MIAPE-CC	蛋白质组学实验的最低信息要求-柱层析	记录柱色谱实验的最小信息集合。
MIAPE-CE	蛋白质组学实验的最低信息要求-毛细管电泳	记录毛细管电泳实验的最小信息集合。
MIAPE-Quant	蛋白质组学实验的最低信息要求-质谱定量	在蛋白质组学实验中使用定量技术所需的最小信息集合，这些信息应足以支持数据的有效解释和评估以及数据资料分析结果的可能的再现性。
MIAPPE	植物表型实验的最低信息要求	植物表型实验报告指南，涵盖植物表型实验以下方面的描述：研究、环境、实验设计、样本管理、生物来源、处理和表型。
MIARE	RNAi实验的最低信息要求	一套描述了关于RNAi实验需要的最小信息集合的指南，以便能够对结果进行明确的解释和重现。
MIASE	模拟实验的最低信息要求	从给定的一组定量模型中获得的，以便能够执行和重现数值模拟实验的所需信息。
MIFlowCyt	流式细胞术实验的最低信息要求	流式细胞术实验概述、样本、仪器和数据分析的信息记录和报告的标准。
MICEE	心脏电生理实验的最低信息要求	由国际领先的实验团队发布的一个最小信息集合，这些信息应足以支撑重现和利用已发表的心脏实验电生理研究。
MIxS - MIGS/ MIMS	（宏）基因组序列的最低信息要求	用于扩展INSDC（DDBJ/EMBL/GenBank）的最小信息集合，以便进一步描述基因组和宏基因组序列。MIMS补充描述了环境方面的必要信息。
MINI	神经科学调查的最低信息要求	神经科学研究中使用电生理学所需的最小信息集合。
MIRIAM	模型注释中所需的最低信息要求	一套用于生物计算模型的一致标注和筛选的指南。该指南应适用于任何结构化格式的计算模型
MISFISHIE	原位杂交和免疫组织化学实验所需的最低信息要求	在发布、公开或交换基于视觉解释的组织基因表达定位实验（如原位杂交、免疫组织化学、报告基因构建遗传实验（GFP/绿色荧光蛋白、 $\beta$ -半乳糖苷酶）等）的结果时，应提供的最小信息集合。
MIxS	任何（X）序列的最低信息要求	序列元数据的总体框架，包括之前MIGS和MIMS标准中的技术特定清单，提供了一种引入MIMARKS等其他清单的方法，并允许使用环境信息对样本数据进行注释。

MIAPepAE	肽阵列实验的最低信息要求	完成肽阵列实验的所需的数据和元数据最小信息集合。
STREND	酶学数据报告标准指南	在酶学实验中正确描述测定条件和酶活性数据所需的最小信息集合。
STROBE	加强流行病学观察性研究的报告	由流行病学家、方法学家、统计学家、研究人员和期刊编辑参与的国际合作倡议创建，旨在加强流行病学观察性研究的报告。
STROBE-nut	加强流行病学观察性研究的报告——营养流行病学	营养流行病学和膳食评估研究报告指南。

### B.3 特定领域的本体、分类法和受控词汇表

#### B.3.1 医学、健康与疾病

表 B.2 医学、健康相关的特定本体、分类法和结构化词汇和疾病

首字母缩写	名字	描述和主页
ATC	解剖学治疗学分类	根据药物作用的器官或系统及其治疗、药理和化学性质对药物活性成分进行分类。它由世界卫生组织药物统计方法合作中心（WHOCC）整理，于1976年首次出版。
CHMO	化学方法本体论	化学实验、材料分析和合成中的数据收集。
CMO	临床测量本体论	用于临床和模式生物研究的形态/生理测量记录。
DCM	DICOM受控术语	DICOM的受控术语。
DINTO	药物相互作用本体论	药物相互作用知识的正规表述形式。
DOID	疾病本体论	按病因学分类的人类疾病描述。
DRON	药物本体论	支持研究索赔数据的比较有效性研究。
ICD-10	国际疾病分类（ICD）第10版	国际疾病分类（ICD）是所有一般流行病学、大部分健康管理和临床使用的国际标准诊断分类。ICD-10是世界卫生组织（WHO）的医学分类列表《国际疾病和相关健康问题统计分类》的第十次修订版。它包含疾病、体征和症状、异常发现、主诉、社会环境以及损伤或疾病的外部原因的代码。
ICD-11	国际疾病分类（ICD）第11版	国际疾病分类（ICD）允许对影响健康的状况和因素进行记录、报告和分类，它包含了疾病、与健康有关的状况和伤害或死亡的外部原因的类别。ICD的目的是允许对在不同国家或地区和不同时间收集的死亡率和发病率数据进行系统的记录、分析、解释和比较。ICD将疾病和其他健康问题的诊断以字母数字代码表示，以便数据的存储、检索和分析。ICD已经成为所有一般流行病学、大部分健康管理和临床使用的国际标准诊断分类。
NCI	NCI叙词表	NCI和其他系统的参考术语表。它涵盖了临床护理，转化和基础研究，以及公共信息和行政活动的词汇。
MedDRA	医药监管活动术语字典	药品监管过程中用于制药行业监管机构的术语，包括从上市前到上市后的活动，以及数据录入、检索、评估和呈现的术语。
MESH	医学主题词	用于索引生命科学期刊文章和书籍的受控词汇表，可用于同义词搜索。
NDFRT	国家药品档案参考术语	描述药物生理作用（PE）的参考层次结构。
OGMS	全科医学本体	涉及临床会诊的实体，包括在医学领域中广泛使用的一般性术语。OGMS的范围限于人类，但



	论	许多术语也可以适用于各种生物体。OGMS提供了一种疾病理论，可以通过具体的疾病本体论进行进一步的详细阐述。
OBIB	生物库本体论	用于生物库存储和管理的注释和建模的本体论。它基于生物医学调查本体论（OBI）的子集，使用基本形式本体论（BFO）作为顶层本体论，并遵循OBO Foundry原则进行开发。本体论的第一个版本合并了两个现有的与生物库相关的本体论，即OMIABIS和生物库本体论。
OMRSE	医学相关社会实体的本体论	涵盖与卫生保健相关的社会实体领域，如人口统计信息（例如，记录个人认同性别（非生理性别）和婚姻状况的社会实体）以及各种个人和组织（患者、医院等）的角色。
OMIABIS	MIABIS本体论	MIABIS（生物库数据共享的最小信息集合）本体论。
RxNORM	RxNORM	临床药物的标准化名称，并将其名称与药店管理和药物相互作用软件中常用的药物词汇相链接。
SNOMED-CT	医学系统命名法-临床术语	系统性组织的、计算机可处理的医学术语集合，提供临床文件和报告中使用的代码、术语、同义词和定义。SNOMED-CT是一个参考术语，可用于跨医疗保健学科的标准化医疗保健语言的交叉映射。
SYMP	症状本体论	疾病症状，包括患者所报告的功能、感觉或外观上的感知变化，这些变化是疾病的指示性症状。
UMLS	统一医学语言系统	统一医学语言系统（UMLS）整合并分发关键术语、分类和编码标准以及相关资源，以促进创建更有效和可互操作的生物医学信息系统和服务，包括电子健康记录。

### B.3.2 解剖学

表 B.3 解剖学相关的特定本体、分类法和结构化词汇

首字母缩写	名字	描述和主页
AEO	解剖实体本体	扩展通用解剖参考本体（CARO）的解剖结构描述。
BSP0	生物空间本体论	空间概念、解剖轴、梯度、区域、平面、侧面和表面的本体论。
CARO	通用解剖学参考本体论	促进不同物种的解剖学本体之间的可操作性。
FMA	解剖学基础模型	关于人体解剖学知识体系的领域本体论。它的本体论框架可以应用和扩展到所有其他物种。
RadLex	放射学词典	放射学相关的实践、教育和研究术语。
UBERON	UBER解剖学本体论	涵盖动物并连接多个物种特异性本体的跨物种解剖学本体。它代表了根据传统解剖学标准（如结构、功能和发育谱系）分类的各种实体。该本体论包括与特定分类的解剖学本体的全面关系，允许功能、表型和表达数据的整合。

### B.3.3 生物化学

表 B.4 生物化学相关的特定本体、分类法和结构化词汇

首字母缩写	名字	描述和主页
CHEBI	生物相关的化学实体	对生物小分子化合物进行结构化分类。
CHEMINF	化学信息本体论	代表化学信息本体论。特别是，它旨在创建一个表示化学结构的本体论，用于丰富地描述化学性质，无论是固有的还是计算得出的。
EMO	酶机制本体论	描述酶的成分及其反应机制的概念，包括这些成分在其中的作用。

## B.3.4 细胞

表 B.5 细胞相关的特定本体、分类和结构化词汇

首字母缩写	名字	描述和主页
CL	细胞本体论	用于表示原核生物、真菌和真核生物细胞类型的本体论。细胞本体论合并了物种特异性解剖本体论中包含的信息，并参考了其他OBO Foundry 本体论，如蛋白质本体论（PR）用于表达独特的生物标志物，基因本体论（GO）用于细胞类型参与的生物过程。
CBO	细胞行为本体论	用于描述多细胞计算模型的术语/实体，重点关注存在性细胞行为（空间性、生长、移动、粘附、死亡等）以及这些行为的计算模型。
CLO	细胞系本体论	用于规范和整合细胞系信息并支持计算机辅助推理的术语。

## B.3.5 基因、蛋白质和RNA

表 B.6 基因、蛋白质和 RNA 相关的特定本体、分类法和结构化词汇

首字母缩写	名字	描述
BioPAX	生物通路交换语言	旨在实现生物通路数据的集成、交换、可视化和分析的标准语言。
GO	基因本体论	结构化词汇表，供研究界用于基因、基因产物和序列的注释。GO定义了用于描述基因功能和这些概念之间关系的概念和类别。
MOP	分子过程本体论	用于描述分子过程的结构化词汇。
OMIT	微小RNA靶标本体论	用于描述微小RNA（miRNA）领域中的数据交流标准和常用数据元素的及结构化词汇。
PW	通路本体论	用于将基因产物注释到通路的结构化词汇。
PPIO	蛋白质相互作用本体论	用于注释与蛋白质相互作用相关实验的结构化词汇。
PRO	蛋白质本体论	用于表示与蛋白质相关的实体的术语/实体。
RNAO	RNA本体论	用于描述RNA所有方面的术语/实体，从 初级序列到比对、二级和三级结构，从碱基配对和碱基堆叠到复杂基序。
SO	序列本体论	用于序列注释、交流注释数据和描述数据库中序列对象的结构化词汇。

## B.3.6 表型

表 B.7 表型相关的特定本体、分类法和结构化词汇

首字母缩写	名字	描述和主页
HP	人类表型本体论	用于描述人类遗传和其他疾病所遇到表型特征的结构化词汇。
MP	哺乳动物表型本体论	用于注释哺乳动物表型数据的标准术语。
NCBITAXON	NCBI分类学	NCBI生物分类学本体论。它不遵循单一来源的分类学规范，而是试图整合来自各种来源的系统发育和分类学知识，包括已发表的文献、基于网络的数据库以及序列提交者和外部分类学专家的建议。
OMIM Ontology	在线人类孟德尔遗传数据库本体论	OMIM是一个全面的、权威的、并且可以免费获取并每日更新的人类基因、遗传表型以及它们之间关系的汇编。OMIM本体包含在OMIM数据库中使用的术语。



OMP	微生物表型本体论	用于注释微生物表型的本体论，包括细菌、古细菌、原生生物、真菌和病毒。
PATO	表型质量本体论	表型质量的本体论，用于许多应用，主要是表型注释。此本体论可与其他本体论（如GO或解剖学本体论）结合使用，以指代表型。
ORDO	罕见病本体论	用于捕捉疾病、基因和其他相关特征之间关系的罕见疾病结构化词汇。

### B.3.7 实验

表 B.8 实验相关的特定本体、分类法和受控词汇

首字母缩写	名字	描述和主页
BAO	生物测定本体论	用于化学生物学筛选测定及其结果的概念，包括高通量筛选（HTS）数据，以便对测定和数据分析进行分类。
ECO	证据与结论本体论	用于描述证据类型和断言方法的术语（类）。ECO术语在生物审编过程中用于获取支持生物结论的证据（例如，基因产物X具有功能Y，由证据Z支持）。获取这些信息可以追溯注释的来源，建立质量控制措施，并对证进行查询。
XCO	实验条件本体论	用于在临床和涉及人类或模式生物的研究中，进行生理和形态测量的条件的结构化词汇。
EFO	实验因素本体论	在多个资源中建模实验变量的概念，以增加当前资源中的注释丰富性和一致性，便于自动注释和集成外部数据。
LOINC	逻辑观测标识符名称和代码	临床和实验室观察的通用语言（标识符、名称和代码的集合）。LOINC是一个测量目录，包括实验室测试、临床测量（如生命体征和人体测量）、标准化调查工具等。LOINC通过提供一组通用代码和结构化名称来交换和聚合临床结果，以明确标识可测量或可观察的事物，用于护理交付、结果管理和研究。
PSI-MI CV	PSI分子相互作用结构化词汇	用于蛋白质相互作用相关实验注释的结构化词汇。由HUPO蛋白质组学标准计划开发。
OBI	生物医学研究本体论	提供一个具有精确定义的含意的术语的开放标准，以描述生物学和医学领域进行调查的所有方面。
SEP	样品处理和分离技术本体论	用于注释科学实验中样本处理和分离技术的结构化词汇，包括但不限于凝胶电泳、柱色谱、毛细管电泳、离心等。
SP	SMART方案	实验方案的结构化词汇，包括用于撰写实验方案所需和足够的信息的概念。

### B.3.8 分析和统计

表 B.9 分析和统计学相关的特定本体、分类法和结构化词汇

首字母缩写	名字	描述和主页
CDAQ	数据比较分析本体论	与进化比较分析相关的概念和关系的结构化词汇。
MAMO	数学建模本体论	对生命科学中使用的数学模型类型、变量、关系和其他相关特征进行分类。
MMO	测量方法本体论	表示用于在临床和模型生物体中进行定性和定量临床和表型测量的各种方法的结构化词汇。
OBCS	生物与临床统计学本体论	生物医学和临床统计学领域的生物医学本体论。OBCS主要用于生物、

		生物医学和临床领域的统计报告。OBCS使用基本形式本体（BFO）作为上层本体。OBCS在生物医学研究本体论（OBI）中引入所有与生物统计学相关的术语，包括所有逻辑公理。
ONTODM-CORE	核心数据挖掘实体本体论	数据挖掘领域的通用本体论。该本体论包括在该领域内所发生的信息处理过程，以及信息处理过程的参与者及其规范。由于OntoDM严格遵守公认的标准，并且与现有的信息处理资源兼容，因此其可移植性和可扩展性较高。其宽泛的涉及范围使得该本体论具有广泛的应用，比如对数据挖掘情景的语义注释，基于本体的QSAR支持等。
STATO	统计方法本体论	用于诸如统计检验等过程的本体论，包括其应用条件以及统计方法所需或产生的信息，如概率分布、变量、扩展和变异度量。STATO还涵盖实验设计的各个方面以及常用于提供数据分布或布局的视觉提示和结果审查的图表和图形表示的描述。

### B.3.9 系统生物学、仿真与虚拟生理人（VPN）

表 B.10 系统生物学、仿真与虚拟生理人相关的特定本体、分类法和结构化词汇

首字母缩写	名字	描述和主页
HUPSON	人体生理仿真本体论	仿真领域的基本概念，用于模拟、模型、算法和该领域中其他资源的通用语义和可操作性。
KISAO	动力学仿真算法本体论	用于在描述仿真实验时引用相关的算法，从算法特征和参数两个方面对现有算法及其相互关系进行了描述和分类。
OPB	生物物理学本体论	应用于生物系统动力学的经典物理学概念。
SBO	系统生物学本体论	用于系统生物学和计算模型中的常用结构化词汇。
TEDDY	描述动力学的术语	系统生物学和合成生物学中生物模型和生物系统的动态行为、可观察的动态现象和控制元素的本体论。

注：有关系统生物学、仿真与虚拟生理人领域的特定语义标准的更多信息，请参阅参考文献和。

### B.3.10 数据类型、格式等

表 B.11 数据类型、格式等相关的特定本体、分类法和结构化词汇

首字母缩写	名字	描述和主页
EDAM	数据和方法本体论	生物信息学中通用的结构化词汇，包括数据类型、数据标识符、数据格式、操作和主题。
IAO	信息产品本体	用于描述信息实体的结构化词汇。
Outdo	数据类型本体论	用于描述数据类型、数据类型生成器、数据类型质量和其他表示任意复杂数据类型的结构化词汇。
RO	OBO关系本体论	多个本体间通用的关系类型。
UCUM	计量单位统一代码	当代国际科学、工程和商业中同时使用的度量单位，其目的是促进数量及其单位的明确电子通信。重点是电子通信，而不是人与人之间的通信。
UO	单位本体	与PATO结合使用的度量单位。

### B.3.11 环境、农业、粮食和营养

表 B.12 农业、食品和营养相关的特定本体、分类法和结构化词汇

首字母缩写	名字	描述和主页
AgrO	农学本体论	在农艺实验中使用的农艺实践、农艺技术和农艺变量，并补充其他作物、牲畜和鱼类本体论，以实现数据收集的统一方法，促进数据共享和重复使用。
ENVO	环境本体论	支持使用环境描述符对任何生物体或生物样本进行注释。这些术语相互关联，通过逻辑公理来描述它们的组成、共定位以及与环境 and 生物过程的关系。使用ENVO术语进行环境描述可以全面描述环境，这对于机器辅助的环境数据集成、存档和联合搜索至关重要。
FAO	宏观真菌解剖学本体论	真菌解剖学。
FoodOn	食物本体论	可作为人类和驯养动物食物的动物、植物和真菌的部分，以及衍生食品及其制作过程。
ONE	营养流行病学本体论	用于描述营养流行病学研究的结构化词汇。
ONS	营养学研究本体论	用于营养学研究的结构化词汇。
PO	植物本体论	将植物解剖学、形态学、生长发育和植物基因组学数据联系起来。

## 参 考 文 献

- [1] Atwood, T.K., Pettifer, S.R., Thorne, D. Bioinformatics challenges at the interface of biology and computer science: mind the gap. UK: John Wiley & Sons, Ltd., 2016
- [2] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., Da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., Van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Monsa, B. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 2016, 3, 160018, pp. 1–9
- [3] Duerst, M., Suigard, M. Internationalized Resource Identifiers (IRIs). The Internet Society, 2005. Available from [viewed 6 April 2022]: <https://datatracker.ietf.org/doc/html/rfc3987#page-3>
- [4] IETF RFC 3986:2005, Uniform Resource Identifier (URI): Generic Syntax
- [5] Berners-Lee, T., Universal Resource Identifiers – Axioms of Web Architecture, 1996. Available from [viewed 6 April 2022]: <http://www.w3.org/DesignIssues/Axioms.html>
- [6] Internet Assigned Numbers Authority (IANA). Available from [viewed 6 April 2022]: [www.iana.org](http://www.iana.org)
- [7] ECMA-404. The JSON Data Interchange Syntax. ECMA International. Second Edition, 2017
- [8] Bray, T., Paoli, J., Sperberg, C.M., Maier, E., Yergeau, F. Extensible Markup Language (XML) 1.0 (Fifth Edition) 2008, W3C Recommendation. Available from [viewed 6 April 2022]: <https://www.w3.org/TR/REC-xml/#sec-origin-goals>
- [9] W3C. Resource Description Framework (RDF) 1.1 Primer. W3C Working Group Note 24, 2014. Available from [viewed 6 April 2022]: <https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/>
- [10] W3C. Web Ontology Language (OWL), 2012. Available from [viewed 6 April 2022]: <https://www.w3.org/OWL/>
- [11] Ciccarese, O., Ocana, M., Castro, L.J.G., Das, S., Clark, T. An open annotation ontology for science on web 3.0. *Journal of Biomedical Semantics*. [online]. 2011, May. 2 Suppl 2: S4. doi: 10.1186/2041-1480-2-S2-S4
- [12] Helmy, M., Crits-Christoph, A., Bader, G.D. Ten Simple Rules for Developing Public Biological Databases. *PLOS Comput. Biol.* 2016, 12(11), pp. 1–8
- [13] Moss, G.P. Recommendations on Biochemical & Organic Nomenclature, Symbols & Terminology etc., International Union of Biochemistry and Molecular Biology, 2020. Available from [viewed 6 April 2022]: <https://iubmb.qmul.ac.uk/>
- [14] W3C. PROV-Overview An Overview of the PROV Family of Documents, 2013. Available from [viewed 6 April 2022]: <https://www.w3.org/TR/prov-overview/>
- [15] Clark, T., Martin, S., Leifeld, T. Globally distributed object identification for biological knowledgebases. 2004, 5(1), pp. 59–70

- [16] Courtot, M., Malone, J., Mungall, C.J. Ten simple rules for biomedical ontology development. *CEUR Workshop Proc.* 2016, 1747
- [17] Courtot, M., Juty, N., Knüpfer, C., Waltemath, D., Zhukova, A., Dräger, A., Dumontier, M., Finney, A., Golebiewski, M., Hastings, J., Hoops, S., Keating, S., Kell, D.B., Kerrien, S., Lawson, J., Lister, A., Lu, J., Machne, R., Mendes, P., Pocock, M., Rodriguez, N., Villegier, A., Wilkinson, D.J., Wimalaratne, S., Laibe, C., Hucka, M., Le Novère, N. Controlled vocabularies and semantics in systems biology. *Mol. Syst. Biol.* 2011, 7:548
- [18] W3C. Data Activity. Available from [viewed 6 April 2022]: <https://www.w3.org/2013/data/>
- [19] OLS. Ontology Search. Available from [viewed 16 May 2022]: [https://www.ebi.ac.uk/ols/ontologies/fma/terms?obo\\_id=FMA:14515](https://www.ebi.ac.uk/ols/ontologies/fma/terms?obo_id=FMA:14515)
- [20] Foundational model of anatomy. Available from [viewed 6 April 2022]: <https://doi.org/10.25504/FAIRsharing.x56jsy>
- [21] The National Center for Biotechnology Information (NCBI). Taxonomy Database. Available from [viewed 6 April 2022]: <https://www.ncbi.nlm.nih.gov/taxonomy>
- [22] Uberon. Available from [viewed 6 April 2022]: <https://uberon.github.io/>
- [23] Natale, D.A., Arighi, C.N., Blake, J.A., Bona, J., Chen, C., Chen, S.C., Christie, K.R., Cowart, J., D'Eustachio, P., Diehl, A.D., Drabkin, H.J., Duncan, W.D., Huang, H., Ren, J., Ross, K., Ruttenberg, A., Shamovsky, V., Smith, B., Wang, Q., Zhang, J., El-Sayed, A., Wu, C.H. Protein Ontology (PRO): enhancing and scaling up the representation of protein entities. *Nucleic Acids Res.* 2017, 45(D1), pp. D339–D346
- [24] Gene Ontology Consortium. The Gene Ontology resource: enriching a Gold mine. *Nucleic Acids Res.* 2021, 49(D1), pp. D325–D334
- [25] Choi, J.Y., Davis, M.J., Newman, A.F., Ragan, M.A. A Semantic Web Ontology for Small Molecules and Their Biological Targets. *J. Chem. Inf. Model.* 2010, 50, pp. 732–741
- [26] Buttigieg, P.L., Morrison, N., Smith, B., Mungall, C.J., Lewis, S.E. The environment ontology: contextualising biological and biomedical entities. *J Biomed Semant.* 2013, 4(1), p. 43
- [27] Bandrowski, A., Brinkman, R., Brochhausen, M., Brush, M.H., Bug, B., Chibucos, M.C., Clancy, K., Courtot, M., Derom, D., Dumontier, M., Fan, L., Fostel, J., Fragoso, G., Gibson, F., Gonzalez-Beltran, A., Haendel, M.A., He, Y., Heiskanen, M., Hernandez-Boussard, T., Jensen, M., Lin, Y., Lister, A.L., Lord, P., Malone, J., Manduchi, E., McGee, M., Morrison, N., Overton, J.A., Parkinson, H., Peters, B., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Schober, D., Smith, B., Soldatova, L.N., Stoeckert Jr C.J., Taylor, C.F., Torniai, C., Turner, J.A., Vita, R., Whetzel, P.L., Zheng, J. The Ontology for Biomedical Investigations. *PLoS One.* 2016, 11(4): e0154556
- [28] US Department of Health and Human Services – National Institutes of Health – National Cancer Institute. NCI Thesaurus. Available from [viewed 6 April 2022]: <https://ncithesaurus.nci.nih.gov/ncitbrowser/>
- [29] Computational Modeling in Biology Network (COMBINE). BioModels.net Qualifiers. Available from [viewed 6 April 2022]: <https://co.mbine.org/standards/qualifiers>
- [30] The OBO Foundry. Available from [viewed 6 April 2022]: <https://www.obofoundry.org/>

- [31] FAIRsharing [online]. Available from [viewed 6 April 2022]: <https://fairsharing.org/>
- [32] Sansone, S.A., McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A.L., Thurston, M., the FAIRsharing Community. FAIRsharing as a community approach to standards, repositories and policies. *Nat. Biotechnol.* 2019, 37(4), pp. 358–367
- [33] FAIRsharing. ISO TC 276 Biotechnology ISO 20691 Collection [online]. Available from [viewed 6 April 2022]: <https://fairsharing.org/3533>
- [34] Needleman, S.B., Wunsch, C.D.A. General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* 1970, 48, pp. 443–453
- [35] Fitch, W.M. An Improved Method of Testing for Evolutionary Homology. *J. Mol. Biol.* 1966, 16, pp. 9–16
- [36] Fitch, W.M. The Relation between Frequencies of Amino Acids and Ordered Trinucleotides. *J. Mol. Biol.* 1966, 16, pp. 1–8
- [37] Lipman, D.J., Pearson, W.R. Rapid and sensitive protein similarity searches. *Science.* 1985, 227, pp. 1435–1441
- [38] Pearson, W.R., Lipman, D.J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA.* 1988, 85, pp. 2444–2448
- [39] US National Institute of Health, National Library of Medicine, National Center for Biotechnology Information. Blast, Blast Topics, Accepted input formats, FASTA. Available from [viewed 6 April 2022]: [https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE\\_TYPE=BlastDocs&DOC\\_TYPE=BlastHelp](https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp)
- [40] Cock, P.J.A, Fields, C.J., Goto, N., Heuer, M.L. Rice, P.M. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. 2010. *Nucleic Acids Res.* 2010, 38(6), pp. 1767–1771
- [41] International Nucleotide Sequence Database Collaboration. Available from [viewed 6 April 2022]: <https://www.insdc.org/>
- [42] Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D. The human genome browser at UCSC. *Genome Res.* 2002, 12, pp. 996–1006
- [43] Kent, W.J., Zweig, A.S., Barber, G., Hinrichs, A.S., Karolchik, D. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics.* 2010, 26(17), pp. 2204–2207
- [44] General Feature Format 3. Available from [viewed 6 April 2022]: <http://gmod.org/wiki/GFF3>
- [45] Generic Model Organism Database. Available from [viewed 6 April 2022]: <http://gmod.org/wiki/Overview>
- [46] The Variant Call Format (VCF) Version 4.2 Specification. Available from [viewed 6 April 2022]: <https://samtools.github.io/hts-specs/VCFv4.2.pdf>
- [47] GTF2. 2: A Gene Annotation Format. Available from [viewed 6 April 2022]: <https://mblab.wustl.edu/GTF22.html>
- [48] Reese, M.G., Moore, B., Batchelor, C., Salas, F., Cunningham, F., Marth, G.T., Stein, L., Flicek, P., Yandell, M., Eilbeck, K. A standard variation file format for human genome sequences. *Genome Biol.* 2010, 11:R88

- [49] Galdzicki, M., Clancy, K. P., Sauro, H. M. The Synthetic Biology Open Language (SBOL) provides a community standard for communicating designs in synthetic biology. *Nat. Biotechnol.* 2014, 32, pp. 545–550
- [50] Hsi-Yang Fritz, M., Leinonen, R., Cochrane, G., Birney, E. Efficient storage of high throughput DNA sequencing data using reference-based compression. *Genome Res.* 2011 May; 21(5):734–40. doi: 10.1101/gr.114819.110
- [51] GA4GH File Encryption Standard. Available from [viewed 6 April 2022]: <http://samtools.github.io/hts-specs/crypt4gh.pdf>
- [52] Thompson, J.D., Higgins, D.G., Gibson, T.J. Clustal W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994, 22(22), pp. 4678–4680
- [53] Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009, 25(16), pp. 2078–2079
- [54] The SAM/BAM Format Specification Working Group. Sequence alignment/Map specification. Available from [viewed 6 April 2022]: <https://samtools.github.io/hts-specs/SAMv1.pdf>
- [55] PHYLIP [online]. Available at: <https://evolution.genetics.washington.edu/phylip/general.html>
- [56] GAME [online]. Available at: <https://www.bioxml.org/Projects/game/game0.1.html>
- [57] Parsons, J.D., Rodriguez-Tomé, P. JESAM: CORBA software components to create and publish EST alignments and clusters. *Bioinformatics.* 2000, 16(4), pp. 313–25
- [58] Waugh, A., Gendron, P., Altman, R., Brown, J.W., Case, D., Gautheret, D., Harvey, S.C., Leontis, N., Westbrook, J., Westhof, E., Zuker, M., Major, F. RNAML: A standard syntax for exchanging RNA information. *RNA.* 2002, 8, pp. 707–717
- [59] Antczak, M., Popenda, M., Zok, T., Zurkowski, M., Adamiak, R.W., Szachniuk, M. New algorithms to represent complex pseudoknotted RNA structures in dot-bracket notation. *Bioinformatics.* 2018, 34(8), pp. 1304–1312
- [60] The Gutell Lab. Comparative RNA Web site and Project. The University of Texas at Austin. Available from [viewed 6 April 2022]: <https://crw-site.chemistry.gatech.edu/>
- [61] Martens, L., Chambers, M., Sturm, M., Kessner, D., Levander, F., Shofstahl, J., Tang, W.H., Römpf, A., Neumann, S., Pizarro, A.D., Montecchi-Palazzi, L., Tasman, N., Coleman, M., Reisinger, F., Souda, P., Hermjakob, H., Binz, P.-A., Deutsch, E.W. mzML—a community standard for mass spectrometry data. *Mol. Cell Proteomics.* 2011, 10(1), R110.000133
- [62] Griss, J., Jones, A.R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., Thallinger, G.G., Reza M. Salek, R.M., Steinbeck, C., Neuhauser, N., Cox, J., Neumann, S., Fan, J., Reisinger, F., Xu, Q.-W., Del Toro, N., Perez-Riverol, Y., Ghali, F., Bandeira, N., Xenarios, I., Kohlbacher, O., Vizcaíno, J.A., Hermjakob, H. The mzTab Data Exchange Format: Communicating Mass-spectrometry-based Proteomics and Metabolomics Experimental Results to a Wider Audience. *Technological Innovation and Resources. Mol. Cell. Proteomics.* 2014, 13(10), pp. 2765–2775
- [63] Gibson, F. et al. The Gel Electrophoresis Markup Language (GelML) from the Proteomics Standards Initiative. *Proteomics.* 2010, 10(17), pp. 3073–3081

- [64] Hermjakob, H. et al. The HUP0 PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol.* 2004, 22(2), pp. 177–183
- [65] Jones, A.R., Pizarro, A., Spellman, P., Miller, M., FuGE Working Group. FuGE: Functional Genomics Experiment Object Model. *OMICS.* 2006, 10(2), pp. 179–184
- [66] Bustin, S.A., Benes, V., Garson, J.A., Hellems, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., Vandesompele, J., Wittwer, C.T. The MIQE Guidelines: Minimum Information for Publication of Quantitative Real-Time PCR Experiments. *Clin. Chem.* 2009, 55(4), pp. 611–622
- [67] Huggett, J.F., Foy, C.A., Benes, V., Emslie, K., Garson, J.A., Haynes, R., Hellems, J., Kubista, M., MUELLER, R.D., Nolan, T., Pfaffl, M.W., Shipley, G.L., Vandesompele, J., Wittwer, C.T., BUSTIN, S.A. The Digital MIQE Guidelines: Minimum Information for Publication of Quantitative Digital PCR Experiments. *Clin. Chem.* 2013, 59(6), pp. 892–902
- [68] Field, D., Sterk, P., Kottmann, R., De Smet, J.W., Amaral-Zettler, L., Cochrane, G., Cole, J.R., Davies, N., Dawyndt, P., Garrity, G.M., Gilbert, J.A., Glöckner, F.O., Hirschman, L., Klenk, H.P., Knight, R., Kyrpides, N., Meyer, F., Karsch-Mizrachi, I., Morrison, N., Robbins, R., San Gil, I., Sansone, S., Schriml, L., Tatusova, T., Ussery, D., Yilmaz, P., White, O., Wooley, J., Caporaso, G. Genomic standards consortium projects. *Stand Genomic Sci.* 2014, 9(3), pp. 599–601
- [69] wwPDB Consortium. Protein Data Bank: the single global archive for 3D macromolecular structure data. 2019, *Nucleic Acids Research*, 47, D520–D528
- [70] Protein Database — PDB. 1992, Atomic Coordinate and Bibliographic Entry Format Description
- [71] Protein Database — PDB. 1996, Protein Data Bank Contents Guide, Atomic Coordinate Entry Format Description, Version 2.2
- [72] Protein Database — PDB. 1996m, Protein Data Bank Contents Guide, Atomic Coordinate Entry Format Description, Version 3.30
- [73] Westbrook, J., Ito, N., Nakamura, N., Henrick, K., Berman, H.M. PDBML: the representation of archival macromolecular structure data in XML. *Bioinformatics.* 2005, 21(7), pp. 988–992
- [74] Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., Pletnev, I. InChI—the worldwide chemical structure identifier standard. *Journal of Cheminformatics.* 2013, 5(7), pp. 1–9
- [75] Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* 1988, 28, pp. 31–36
- [76] Rayner, T.F., Rocca-Serra, P., Spellman, P.T., Causton, H.C. Farne, A., Holloway, E., Irizarry, R.A., Liu, J., Maier, D.S., Miller, M., Petersen, K., Quackenbush, J., Sherlock, G., Stoeckert Jr, C.J., White, J., Whetzel, P.L., Wymore, F., Parkinson, H., Sarkans, U., Ball, C.A., Brazma, A. A simple spreadsheet-based, MIAME-supportive format for microarray data: MAGE-TAB. *BMC Bioinformatics.* 2006, 7, p. 489
- [77] Brazma, A. et al. Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nat. Genet.* 2002, 29, pp. 365–371



- [78] FAIRsharing.org. ISA-Tab; Investigation Study Assay Tabular (ISA-Tab). Available from [viewed 6 April 2022]: <https://doi.org/10.25504/FAIRsharing.53gp75>
- [79] González-Beltrán, A., Maguire, E., Sansone, S.A., Rocca-Serra, P. linkedISA: semantic representation of ISA-Tab experimental metadata. *BMC Bioinformatics*. 2014, 15(Suppl 14), S4
- [80] Rocca-Serra, P., Sansone, S.A., Brandiz, i M. Specification documentation: ISA-TAB 1.0. Zenodo 2009. Available from [viewed 6 April 2022]: <https://doi.org/10.5281/zenodo.161355>
- [81] FAIRsharing.org. Investigation Study Assay JSON (ISA-JSON). Available from [viewed 6 April 2022]: <https://doi.org/10.25504/FAIRsharing.yhLgTV>
- [82] Johnson, D., Batista, D., Cochrane, K., Davey, R.P., Etuk, A., Gonzalez-Beltran, A., Haug, K., Izzo, M., Larralde, M., Lawson, T.N., Minotto, A., Moreno, P., Nainala, V.C., O'Donovan, C., Pireddu, L., Roger, P., Shaw, F., Steinbeck, C., Weber, R.J.M., Sansone, S.A., Rocca-Serra, P. ISA API: An open platform for interoperable life science experimental metadata. *Gigascience*. 2021, 10(9), giab060
- [83] FAIRsharing.org. Fast Healthcare Interoperability Resources (FHIR). Available from [viewed 6 April 2022]: <https://doi.org/10.25504/FAIRsharing.25k4yp>
- [84] Health Level Seven International (HL7). Available from [viewed 6 April 2022]: <https://www.hl7.org>
- [85] Burel, J.-M., et al. Publishing and sharing multi-dimensional image data with OMERO. *Mamm. Genome*. 2015, 26, pp. 441–447
- [86] Golebiewski, M. Data Formats for Systems Biology and Quantitative Modeling., *Encyclopedia of Bioinformatics and Computational Biology*. 2019, 2, pp. 884–893
- [87] Nickerson, D., Atalag, K., de Bono, B., Geiger, J., Goble, C., Hollmann, S., Lonien, J., Müller, W., Regierer, B., Stanford, N.J., Golebiewski, M., Hunter, P. The Human Physiome: how standards, software and innovative service infrastructures are providing the building blocks to make it achievable. *Interface Focus*. 2016, 6(2), p. 20150103
- [88] Neal, M.L., König, M., Nickerson, D., Mısırlı, G., Kalbasi, R., Dräger, A., Atalag, K., Chelliah, V., Cooling, M.T., Cook, D.L., Crook, S., de Alba, M., Friedman, S.H., Garny, A., Gennari, J.H., Gleeson, P., Golebiewski, M., Hucka, M., Juty, N., Myers, C., Olivier, B.G., Sauro, H.M., Scharm, M., Snoep, J.L., Touré, V., Wipat, A., Wolkenhauer, O., Waltemath, D. Harmonizing semantic annotations for computational models in biology. *Brief. Bioinform*. 2019, 20(2), pp. 540–550
- [89] Keating, S.M., Waltemath, D., König, M., Zhang, F., Dräger, A., Chaouiya, C., Bergmann, E.T., Finney, A., Gillespie, C.S., Helikar, T., Hoops, S., Malik-Sheriff, R.S., Moodie, S.L., Moraru, I.I., Myers, C.J., Naldi, A., Olivier, B. G., Sahle, S., Schaff, J.C., Smith, L.P. et al. SBML Level 3: an extensible format for the exchange and reuse of biological models. *Molecular Systems Biology*. 2020, 16(8), e9110
- [90] Gleeson, P., Crook, S., Cannon, R.C., Hines, M.L., Billings, G.O. et al. NeuroML: A Language for Describing Data Driven Models of Neurons and Networks with a High Degree of Biological Detail. *PLOS Comput. Biol*. 2010, 6(6), p. e1000815
- [91] Britten, R.D., Christie, G.R., Little, C., Miller, A.K., Bradley, C., Wu, A, Yu, T, Hunter, P, Nielsen, P. FieldML, a proposed open standard for the Physiome project for mathematical model representation. *Med. Biol. Eng. Comput*. 2013, 51(11), pp. 1191–1207

- [92] Demir, E., Cary, M., Paley, S. et al. The BioPAX community standard for pathway data sharing. *Nat. Biotechnol.* 2010, 28, pp. 935–942
- [93] Le Novère, N., Hucka, M., Mi, H., Moodie, S., Schreiber, F., Sorokin, A., Demir, E., Wegner, K., Aladjem, M.I., Wimalaratne, S.M., Bergman, F.T., Gauges, R., Ghazal, P., Kawaji, H., Li, L., Matsuoka, Y., Villéger, A., Boyd, S.E., Calzone, L., Courtot, M., Dogrusoz, U., Freeman, T.C., Funahashi, A., Ghosh, S., Jouraku, A., Kim, S., Kolpakov, F., Luna, A., Sahle, S., Schmidt, E., Watterson, S., Wu, G., Goryanin, I., Kell, D.B., Sander, C., Sauro, H., Snoep, J.L., Kohn, K., Kitano, H. The Systems Biology Graphical Notation. *Nat. Biotechnol.* 2009, 27, pp. 735–741
- [94] Waltemath, D., Adams, R., Bergmann, F.T., Hucka, M., Kolpakov, F., Miller, A.K., Moraru, I.I., Nickerson, D., Sahle, S., Snoep, J.L., Le Novère, N. Reproducible computational biology experiments with SED-ML – the Simulation Experiment Description Markup Language. *BMC Syst. Biol.* 2011, 5, p. 198
- [95] Bergmann, F.T., Adams, R., Moodie, S., Cooper, J., Glont, M., Golebiewski, M., Hucka, M., Laibe, C., Miller, A.K., Nickerson, D.P., Olivier, B.G., Rodriguez, N., Sauro, H.M., Scharm, M., Soiland-Reyes, S., Waltemath, D., Yvon, F., Le Novère, N. COMBINE archive and OMEX format: one file to share all information to reproduce a modelling project. *BMC Bioinformatics.* 2014, 15, p. 369
- [96] National Center for Biomedical Ontology (NCBO). NCBO BioPortal. Available from [viewed 6 April 2022]: <https://bioportal.bioontology.org>
- [97] European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI). Ontology Lookup Service (OLS). Available from [viewed 6 April 2022]: <https://www.ebi.ac.uk/ols>
- [98] Wolstencroft, K., Owen, S., Krebs, O., Nguyen, Q., Stanford, N.J., Golebiewski, M., Weidemann, A., Bittkowski, M., An, L., Shockley, D., Snoep, J.L., Mueller, W., Goble, C. SEEK: a systems biology data and model management platform. *BMC Syst Biol.* 2015, 9:33
- [99] Wolstencroft, K., Krebs, O., Snoep, J.L., Stanford, N.J., Bacall, F., Golebiewski, M., Kuzyakiv, R., Nguyen, Q., Owen, S., Soiland-Reyes, S., Straszewski, J., van Niekerk, D.D., Williams, A.R., Malmström, L., Rinn, B., Müller, W., Goble, C. FAIRDOMHub: a repository and collaboration environment for sharing systems biology research. *Nucleic Acids Res.* 2017, 45(D1), pp. D404–D407
- [100] Jacobs, I. Architecture of the World Wide Web Volume One. W3C Recommendation December 2004. Available from [viewed 10 February 2021]: <http://www.w3.org/TR/webarch/>
- [101] W3C Semantic Web. 2001 Available from [viewed 6 April 2022]: <https://www.w3.org/2001/sw/wiki/OWL>
- [102] ISO 1087:2019, Terminology work and terminology science — Vocabulary
- [103] ISO/TR 3985, Biotechnology — Data publication — Preliminary considerations and concepts
- [104] ISO 5127:2017, Information and documentation — Foundation and vocabulary
- [105] ISO 9000:2015, Quality management systems — Fundamentals and vocabulary
- [106] ISO 11799:2015, Information and documentation — Document storage requirements for archive and library materials

- [107] ISO 12052, Health informatics — Digital imaging and communication in medicine (DICOM) including workflow and data management
- [108] ISO 12639, Graphic technology — Prepress digital data exchange — Tag image file format for image technology (TIFF/IT)
- [109] ISO 14199:2015, Health informatics — Information models — Biomedical Research Integrated Domain Group (BRIDG) Model
- [110] ISO 16684-1:2019, Graphic technology — Extensible metadata platform (XMP) — Part 1: Data model, serialization and core properties
- [111] ISO/TS 23029:2020, Web-service-based application programming interface (WAPI) in financial services
- [112] ISO 23903:2021, Health informatics — Interoperability and integration reference architecture — Model and framework
- [113] ISO 24619:2011, Language resource management — Persistent identification and sustainable access (PISA)
- [114] ISO/TS 27790:2009, Health informatics — Document registry framework
- [115] ISO/IEC Guide 99:2007, International vocabulary of metrology — Basic and general concepts and associated terms (VIM)
- [116] ISO/IEC 646, Information technology — ISO 7-bit coded character set for information interchange
- [117] ISO/IEC 2382:2015, Information technology — Vocabulary
- [118] ISO/IEC 5218, Information technology — Codes for the representation of human sexes
- [119] ISO/IEC 8859 (all parts), Information technology — 8-bit single-byte coded graphic character sets
- [120] ISO/IEC TR 10032, Information technology — Reference Model of Data Management
- [121] ISO/IEC 10646, Information technology — Universal coded character set (UCS)
- [122] ISO/IEC 10918-1, Information technology — Digital compression and coding of continuous-tone still images: Requirements and guidelines
- [123] ISO/IEC 11179-1:2015, Information technology — Metadata registries (MDR) — Part 1: Framework
- [124] ISO/IEC 11179-7:2019, Information technology — Metadata registries (MDR) — Part 7: Metamodel for data set registration
- [125] ISO/IEC 11404, Information technology — General-Purpose Datatypes (GPD)
- [126] ISO/IEC 12785-1:2009, Information technology — Learning, education, and training — Content packaging — Part 1: Information model
- [127] ISO/IEC 13250-2:2006, Information technology — Topic Maps — Part 2: Data model
- [128] ISO/IEC 13818-3:1998, Information technology — Generic coding of moving pictures and associated audio information — Part 3: Audio
- [129] ISO/IEC 14957, Information technology — Representation of data element values — Notation of the format
- [130] ISO/IEC 15944-1:2011, Information technology — Business operational view — Part 1: Operational aspects of open-edl for implementation

- [131] ISO/IEC 15948, Information technology — Computer graphics and image processing — Portable Network Graphics (PNG): Functional specification
- [132] ISO/IEC 19502, Information technology — Meta Object Facility (MOF)
- [133] ISO/IEC 20944-1:2013, Information technology — Metadata Registries Interoperability and Bindings (MDR-IB) — Part 1: Framework, common vocabulary, and common provisions for conformance
- [134] ISO/IEC 23092-1, Information technology — Genomic information representation — Part 1: Transport and storage of genomic information
- [135] ISO/IEC 23092-2, Information technology — Genomic information representation — Part 2: Coding of genomic information
- [136] ISO/IEC 23092-3, Information technology — Genomic information representation — Part 3: Metadata and application programming interfaces (APIs)
- [137] ISO/IEC 23092-4, Information technology — Genomic information representation — Part 4: Reference software
- [138] ISO/IEC 23092-5, Information technology — Genomic information representation — Part 5: Conformance
- [139] ISO/IEC 23092-64), Information technology — Genomic information representation — Part 6: Coding of genomic annotations
- [140] ISO/IEC/IEEE 15939:2017, Systems and software engineering — Measurement process
- [141] IEEE 754, IEEE Standard for Floating-Point Arithmetic
-