INTERNATIONAL STANDARD

ISO 20691

First edition
2022-11

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

Biotechnologie — Exigences relatives au formatage et à la description des données dans les sciences de la vie

Reference number
ISO 20691:2022(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 276, *Biotechnology*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

Life science research and the application of the obtained results in the biotechnology, diagnostics and pharmaceutical industries depend on complex data obtained from a wide range of assays, biological and functional studies, as well as process descriptions, laboratory and field measurements. This includes the use of the derived data for computational reconstruction, modelling and simulation of biological, biotechnological and physiological processes, as well as their applications in biotechnological workflows. Data enabled life sciences and biotechnology research span across a wide range of biological and biotechnological domains and applications (e.g. human health, genetically engineered organisms, environmental sciences, agriculture, bioremediation, DNA sequencing, chromatography, microscopy). Data driven, data intensive and big data analytical approaches in the life sciences are possible only with the use of computational methods and through consistent description, structuring and integration of data.[1] Data storage, representation, meaning, interpretation, exchange and re-use are all affected by format design. This document satisfies a critical need to set a framework for interoperable and unambiguous data recording, description and transfer by setting fundamental requirements for data recorded, processed, re-used and exchanged in the life sciences enabling the maximum data value and utilization.

These life science data from different sources and recorded at different times must be findable, accessible, interoperable and reusable (F-A-I-R).[2] Data sets are valuable and useful only if they are accessible and stored in well structured, consistent formats. Data versioning, data archiving and tracing data provenance are ensured by timeless and platform independent formats. Complete and updatable metadata (i.e. data describing the data) facilitates locating, use and analysis of data.

This document provides requirements and recommendations for standardized interoperable life science data formats. It provides a conceptual framework for, as well as references to, many different subdomain-specific data formatting and description standards defined by the biotechnological and biological domain communities. A technology-independent framework of minimal requirements and rules for the coherent utilization of the referenced domain-specific formatting and description standards and their concerted interplay is described. This document, therefore, provides rules and guidelines for coherent, subdomain overarching data formatting and description, as a foundation for data integration across domains. Moreover, rules and guidelines for the creation of (sub-)domain specific standards, their interoperability and their implementations are provided.

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

1 Scope

This document specifies requirements for the consistent formatting and documentation of data and corresponding metadata (i.e. data describing the data and its context) in the life sciences, including biotechnology, and biomedical, as well as non-human biological research and development. It provides guidance on rendering data in the life sciences findable, accessible, interoperable and reusable (F-A-I-R).

This document is applicable to manual or computational workflows that systematically capture, record or integrate data and corresponding metadata in the life sciences for other purposes.

This document provides formatting requirements for both primary experimental or procedural data obtained manually and machine derived data. This document also describes requirements for storing, sharing, accessing, interoperability and reuse of data and corresponding metadata in the life sciences.

This document specifies requirements for large quantities of data systematically obtained from automated high throughput workflows in the life sciences, as well as requirements for large-scale and small-scale data sets obtained by other life science technologies and manual data capture.

This document is applicable to many domains in biotechnology and the life sciences including, but not limited to: basic/applied research in all domains of the life sciences, and industrial, medical, agricultural, or environmental biotechnology (excluding for diagnostic or therapeutic purposes), as well as methodology-driven domains, such as genomics (including massive parallel sequencing, metagenomics, epigenomics and functional genomics), transcriptomics, translatomics, proteomics, metabolomics, lipidomics, glycomics, enzymology, immunochemistry, synthetic biology, systems biology, systems medicine and related fields.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

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 Terms and definitions

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

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
ASCII
American Standard Code for Information Interchange
character encoding standard for electronic communication

Note 1 to entry: ASCII codes represent text in computers, telecommunications equipment and other devices.

Note 2 to entry: Most modern character-encoding schemes are based on ASCII, although they support many additional characters. In an ASCII file, each alphabetic, numeric or special character is represented with a 7-bit binary number (a string of seven 0s or 1s). 128 possible characters are defined.

Note 3 to entry: The 7-bit ASCII is documented in ISO/IEC 646.

3.2
backward compatibility
compatibility of a newer coding standard with an older coding standard where the decoders designed to operate with the older coding standard can continue to operate by decoding all or parts of a bitstream produced according to the newer coding standard

3.3
character
printable symbol having phonetic or pictographic meaning and usually forming part of a word of text, depicting a numeral or expressing grammatical punctuation

3.4
characteristic
abstraction that qualifies a *property* (3.37) of an *object* (3.31) or of a set of objects

[SOURCE: ISO 1087:2019, 3.2.1, modified — "that qualifies a property of an object or of a set of objects" has replaced "of a property", and the example and note to entry have been deleted.]

3.5
class
description of a set of *objects* (3.31) that share the same properties, operations, methods, relationships and semantics

3.6
code
system of rule(s) to convert information such as text, images, sounds or electric, photonic or magnetic signals into another form or representation to facilitate analysis, communication or storage in a storage medium

3.7
concept
unit of knowledge created by a unique combination of *characteristics* (3.4)

[SOURCE: ISO 1087:2019, 3.2.7, modified — Notes 1 and 2 to entry have been deleted.]

3.8
context
circumstance, purpose and perspective under which an *object* (3.31) is defined or used

[SOURCE: ISO/IEC 11179-1:2015, 3.3.7, modified — Note 1 to entry had been deleted.]

3.9
data
reinterpretable representation of information in a formalized manner suitable for communication, interpretation or processing

[SOURCE: ISO/IEC 2382:2015, 2121272, modified — Note 1, 2, and 3 to entry have been deleted.]

3.10
data element
unit of *data* (3.9) that is considered in *context* (3.8) to be indivisible

Note 1 to entry: This term is meant for the organization of data.

Note 2 to entry: The definition states that a data element is "indivisible" in some context. This means it is possible that a data element considered indivisible in one context (e.g. telephone number) can be divisible in another context (e.g. country code, area code, local number).

[SOURCE: ISO/IEC 15944-1:2011, 3.16, modified — "(in organization of data)" was deleted from the term, the example and Note 1 to entry were deleted, and new Notes 1 and 2 to entry were added.]

3.11
data format
arrangement of *data* (3.9) in a file or stream

[SOURCE: ISO/TS 27790:2009, 3.18]

3.12
data integrity
property (3.37) that *data* (3.9) have not been altered or destroyed in an unauthorized manner

[SOURCE: ISO/TS 27790:2009, 3.19]

3.13
data model
graphical and/or lexical representation of *data* (3.9), specifying their properties, structure and interrelationships

[SOURCE: ISO/IEC 11179-1:2015, 3.2.7]

3.14
data provider
individual or organization that is a source of *data* (3.9)

[SOURCE: ISO/IEC/IEEE 15939:2017, 3.5]

3.15
data set
dataset
identifiable collection of *data* (3.9) available for access or download in one or more *data formats* (3.11)

Note 1 to entry: A data set can be a smaller grouping of data, which, though limited by some constraint such as spatial extent or feature type, is located physically within a larger data set. Theoretically, a data set can be as small as a single feature or feature attribute contained within a larger data set.

Note 2 to entry: A data set may be presented in a tabular form and stored and distributed in tables in word processed documents, spread sheets or databases. It could also be presented in any one of a number of alternative formats, including AVRO, JSON, RDF and XML.

[SOURCE: ISO/IEC 11179-7:2019, 3.1.4]

3.16
data type
classification of *data* (3.9) indicating how it can be used

Note 1 to entry: Data type provides a set of values from which an expression can take its values.

Note 2 to entry: It characterizes both the content and the structure of an element.

Note 3 to entry: It characterizes properties of those values and operations on those values.

Note 4 to entry: Data types can be categorized in many ways, e.g. as master data or reference data.

3.17
data representation paradigm
tool for *data* (3.9) representation providing a well-defined syntax that is devoid of any application-level semantics

3.18
entity
any concrete or abstract thing that exists, did exist or can exist, including its properties and interactions with other things

3.19
extensibility
provisions in an early version of a *data format* (3.11) that are designed to maximize the interworking of implementations of that early version with the expected implementations of a later version of that data format

3.20
forward compatibility
compatibility of an older coding standard with a newer coding standard where the decoders designed to operate with the newer coding standard can decode bitstreams of the older coding standard

[SOURCE: ISO/IEC 13818-3:1998, 2.1.108, modified — "compatibility of an older coding standard with a newer coding standard where the" has replaced "A newer coding standard is forward compatible with an older coding standard if".]

3.21
identifier
sequence of *characters* (3.3), capable of uniquely identifying that with which it is associated, within a specified *context* (3.8)

[SOURCE: ISO/IEC 11179-1:2015, 3.1.3, modified — Notes 1 and 2 to entry have been deleted.]

3.22
interoperability
ability of two or more systems or components to exchange information and to use the information that has been exchanged

[SOURCE: ISO/TS 27790:2009, 3.39]

3.23
IRI
internationalized resource identifier
sequence of *characters* (3.3) from the *universal coded character set* (3.51), capable of uniquely identifying that with which it is associated, within a specified *context* (3.8)

Note 1 to entry: 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]

3.24
JSON
JavaScript Object Notation
open and text-based exchange format

Note 1 to entry: Data transmitted in JSON formats make it easy to read and write (for humans), parse and generate (for computers).

[SOURCE: ISO/TS 23029:2020, 3.3]

3.25
long-term storage
storage, for a period of undefined length, of *data* (3.9) kept for permanent retention

[SOURCE: ISO 11799:2015, 2.3, modified — "data" has replaced "material" in the definition.]

3.26
maintainer
maintenance organization
individual or organization that maintains the *data format* (3.11)

3.27
metadata
data (3.9) that defines and describes other *data* (3.9)

[SOURCE: ISO/IEC 11179-1:2015, 3.2.16]

3.28
metadata object
object (3.31) type defined by a metamodel

[SOURCE: ISO/IEC 11179-1:2015, 3.2.18]

3.29
metadata attribute
attribute of an instance of a *metadata object* (3.28) commonly needed in its specification

3.30
namespace
class (3.5) of elements that are used to identify and refer to *objects* (3.31) of various kinds that can be instantiated as *uniform resource identifiers* (3.49)

Note 1 to entry: A namespace ensures that all of a given set of objects have unique names so that they can be easily identified.

Note 2 to entry: Namespaces are commonly structured as hierarchies to allow reuse of names in different contexts.

3.31
object
anything perceivable or conceivable

Note 1 to entry: Objects can be material (e.g. “engine”, “sheet of paper”, “diamond”), immaterial (e.g. “conversion ratio”, “project plan”) or imagined (e.g. “unicorn”, “scientific hypothesis”).

[SOURCE: ISO 1087:2019, 3.1.1]

3.32
ontology
collection of *terms* (3.47), relational expressions and associated natural-language definitions together with one or more formal theories designed to capture the intended interpretations of these definitions

Note 1 to entry: An ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application.

[SOURCE: ISO 23903:2021, 3.18]

3.33
OWL
web ontology language
web-based language designed for use in applications that need to process the content of information

[SOURCE: ISO 14199:2015, 3.6]

3.34
permissible value
designation of a value meaning

[SOURCE: ISO/IEC 11179-1:2015, 3.3.20, modified — Notes 1 and 2 to entry have been deleted.]

3.35
persistent identifier
PID
unique *identifier* (3.21) that ensures permanent access for a digital *object* (3.31) by providing access to it independently of its physical location or current ownership

[SOURCE: ISO 24619:2011, 3.2.4, modified — Note 1 to entry has been deleted.]

3.36
predicate
qualifier
relationship between a *data set* (3.15), or a *data element* (3.10), and a specific subject of a referenced resource

3.37
property
characteristic (3.4) common to all members of a *class* (3.5)

3.38
proprietary software
non-free computer software for which the software's publisher or another person retains intellectual property rights, usually copyright of the source *code* (3.6) and sometimes also patent rights

3.39
provenance
information on the place and time of origin, derivation or generation of a resource or a record or proof of authenticity or of past ownership

[SOURCE: ISO/IEC 11179-7:2019, 3.1.10]

3.40
publisher
individual or organization that has published a *data format* (3.11)

3.41
quantity
quantitative value
property (3.37) of a phenomenon, body, or substance where the property has a magnitude that can be expressed as a number and a reference

[SOURCE: ISO/IEC Guide 99:2007, 1.1, modified — "quantitative value" has been added as the admitted term, and the notes to entry and the example have been deleted.]

3.42
RDF
Resource Description Framework
XML (3.53) syntax for describing metadata

[SOURCE: ISO 16684-1:2019, 3.6]

3.43
reification
making a topic represent the subject of another topic map construct in the same topic map

[SOURCE: ISO/IEC 13250-2:2006, 3.11]

3.44
repository
data repository
implementation of a collection of *data* (3.9) along with data access and control mechanisms, such as search, indexing, storage, retrieval and security

Note 1 to entry: A repository can cover aspects of data governance, data stewardship and data ownership.

[SOURCE: ISO/IEC 20944-1:2013, 3.21.12.19, modified — "repository" has been added as the preferred term and the example has been deleted. Note 1 has been added.]

3.45
semantic interoperability
ability for *data* (3.9) shared by systems to be understood at the level of formally defined domain *concepts* (3.7)

[SOURCE: ISO/TS 27790:2009, 3.67]

3.46
stable format
stable data format
data format (3.11) specification not subject to constant or major changes over time

3.47
term
designation that represents a general *concept* (3.7) by linguistic means

[SOURCE: ISO 1087:2019, 3.4.2, modified — The example and note to entry have been deleted.]

3.48
terminology
set of *terms* (3.47) representing a system of *concepts* (3.7) within a specified domain

Note 1 to entry: This implies a published purpose and scope from which one can determine the degree to which this representation adequately covers the domain specified.

[SOURCE: ISO 1087:2019, 3.1.11, modified — "terms representing a system of concepts within a specified domain" has replaced "designations and concepts belonging to one domain or subject", and Note 1 to entry has been added.]

3.49
URI
uniform resource identifier
compact sequence of *characters* (3.3) that uniquely identifies an abstract or physical resource

Note 1 to entry: See IETF RFC 3986:2005.

[SOURCE: ISO/IEC 12785-1:2009, 3.23, modified — "uniquely" was added to the definition.]

3.50
unit of measure
actual units in which the associated values are measured

Note 1 to entry: The dimensionality of the associated conceptual domain must be appropriate for the specified unit of measure.

[SOURCE: ISO/IEC 11179-1:2015, 3.3.29]

3.51
UCS
universal coded character set
character (3.3) set encoding standard for international electronic communication

3.52
verification
confirmation, through the provision of objective evidence, that specified requirements have been fulfilled

Note 1 to entry: The objective evidence needed for a verification can be the result of an inspection or of other forms of determination such as performing alternative calculations or reviewing documents.

[SOURCE: ISO 9000:2015, 3.8.12, modified — Notes 2 and 3 to entry have been deleted.]

3.53
XML
extensible markup language
markup language that encodes information in a way that is machine-processable as well as human-readable

[SOURCE: ISO 5127:2017, 3.1.9.19]

4 Recommendations and requirements for the description of entities and concepts in life science data

4.1 General

This clause is focused on recommendations and requirements for the consistent description of biological or conceptual entities in life science data and data types (see ISO/IEC 11404) and the usage of ubiquitous persistent identifiers (PIDs) to unambiguously refer to them.

Any biological or conceptual entity or defined process comprised in a data set, corresponding metadata set or data collection shall be made unambiguously identifiable. To this end, persistent entity identifiers in the form of uniform resource identifiers (URIs)[4] or internationalized resource identifiers (IRIs) should be applied for the attribution of a biological or conceptual entity or defined process to the corresponding unambiguous definition or reference of the entity or process. This should be achieved by annotating an entity or process in the data set by using a corresponding URI or IRI to an entry in a database, registry, terminology resource, ontology or other appropriate resource carrying the respective definition or data entry for disambiguation of the entity or process.

4.2 Recommended ubiquitous identifier scheme for biological and conceptual entities

4.2.1 URI provisions

4.2.1.1 General

A biological or conceptual entity or a defined process in a data set, corresponding metadata set or data collection can be represented as or annotated by a URI.[4] A biological or conceptual entity identifier is qualified if it possesses a specific namespace and context, e.g. if it resides in a database, or if it is included in a specific reference.[5] A URI for a biological or conceptual entity can be represented in any appropriate compatible scheme, e.g. http, https, urn or similar. The used URI scheme should be registered with the Internet Assigned Numbers Authority (IANA),[6] although non-registered schemes are also valid. A URI shall be a string of ASCII characters with its format as follows:

scheme://authority/path/name

where “authority” and “path” define the type of data (namespace), i.e. the collection of all “names” of the same type, and “name” refers to the respective biological or conceptual entity within this namespace. “Authority” shall at least comprise the host, consisting of either a registered name or an IP address the namespace refers to (e.g. the database or web resource that carries the namespace or points to it), and “path” comprises at least one or more, hierarchically structured namespace qualifier(s) pointing to a collection of names of the same type. Hierarchical levels within the path shall be defined by forward

slashes ("/"). Of the ASCII character set, the characters: / ? # [ ] @ are reserved for use as delimiters of the generic URI components and shall be percent-encoded ("escaped"), e.g. "%3F" for a question mark. [114]

Dereferencing a URI shall lead to a representation of the distinct biological entity or concept identified by the URI. Two URIs are the same if the escaped version of both URIs are the same, character for character. URIs that are different can be equivalent, but have to be canonicalized by a software agent.

4.2.1.2 Persistence of URIs

Any URI used to describe the data or any of its contained entities or both shall be persistent and shall not change. Provenance and versioning shall be maintained for changes to the data represented by the URI.

4.2.1.3 Metadata for a URI

Any metadata connected to a URI shall be capturable and shall be kept over the whole lifetime of the data.

A URI shall be persistent and remain independent of its mapping on a server, and its notation (including upper versus lower case letters). Although schemes are case-insensitive, the canonical form is lower case for documents that specify schemes. Implementations can accept upper case letters as equivalent to lower case in scheme names (e.g. allow "HTTP" as well as "http") for the sake of robustness.

A URI should not attempt to infer the properties of the biological entity or concept.

A URI shall identify only one biological entity or concept. Using the same URI to identify more than one biological entity or concept, causes URI collision. URI collision shall be avoided. Communities of databases are responsible for avoiding the assignment of equivalent URIs to multiple biological entities or concepts. Communities of databases are responsible for representation management of URIs.

A URI shall be opaque and shall not contain:

a) the author's name;

b) the subject;

c) the status;

d) the access;

e) the file name extension;

f) the software mechanism(s);

g) the disk name;

h) the domain name.

4.2.2 IRI provisions

A biological or conceptual entity or a defined process in a data set, corresponding metadata set or data collection can be represented as or annotated by an IRI.[3] The IRI is a complement to URI. It extends the syntax of URIs to a much wider character set and defines "internationalized" versions corresponding to other constructs, such as URI references.

The IRI shall be used for all entities that do not only use ASCII characters. All other URI provisions from 4.2.1 apply correspondingly also for IRIs.

The IRI shall be represented as a character sequence from the UCS, the universal coded character set (Unicode in accordance with ISO/IEC 10646). IRIs shall be a string of characters from the UCS with format as follows:

scheme://authority/path/name

where "authority" and "path" define the type of data (namespace), i.e. the collection of all "names" of the same type, and "name" refers to the respective biological or conceptual entity within this namespace. Hierarchical levels within the path shall be defined by forward slashes.

4.2.3 Relationship between URI and IRI

Although the ASCII character set is included in UCS at the human readable level, the converse is not true. Therefore, a mapping step can be required to retrieve URIs as IRIs.

IRIs shall require a translation step to facilitate conversion to URIs.[3]

4.3 Formatting data and contextual descriptive data (metadata) for biological entities and concepts

4.3.1 General

Data and metadata formats for biological entities and concepts can vary with community, discipline, institution, nationality and time. This document ensures that the impact of changes to the usability of data and metadata between and amongst communities, disciplines, institutions and nationalities and over time is minimized (see ISO/IEC 14957 and ISO/IEC TR 10032).

4.3.2 Version control

All aspects of data and metadata formats including schemas and contained elements for biological entities, as well as underlying concepts shall be version controlled. Version control for biological entities and concepts shall be achieved with the use of a version control system. The elements of data formats shall be versioned using an established ontology (see Annex B).

4.3.3 Arbitrary Limits

Formats for biological entities and concepts shall not contain arbitrary limits, such as maximum lengths for strings of characters or enforcing upper and lowercase restrictions or distinctions.

4.3.4 Character sets

Formats for biological entities and concepts shall support both ASCII and non-ASCII (i.e. UCS) characters in all languages.

4.3.5 Machine readability

Formats for describing and encoding biological entities and concepts, as well as their corresponding metadata, shall ensure machine readability, and may permit human readability (see ISO/TR 3985).

If formats for describing and encoding biological entities and concepts, as well as their corresponding metadata, are created from manually produced plain text, it is important to prevent breaking changes.

Only generally recognized data representation paradigms, such as JavaScript Object Notation (JSON) [7], extensible markup language (XML)[8] and Resource Description Framework (RDF)[9] or similar concepts, and appropriate domain-specific metadata standards and formats, as well as recognized terminologies (see Annex B for recommended metadata standards and domain-specific terminologies) shall be used to construct data and metadata formats for describing and encoding biological entities and concepts, as well as their corresponding metadata. For the consistent data representation and

structuring, only recognized domain-specific data formats (see Annex A) and metadata models (see Reference [10] and ISO/IEC 19502) shall be used where applicable.

Data and metadata formats that are not open and do not protect against the loss of semantic context during processing or transferring (i.e. in databases) shall be made machine readable subject to:

a) security considerations;

b) cost(s) and benefit(s);

c) legal liabilities;

d) intellectual property right(s);

e) confidential business information;

f) contract restriction(s);

g) other binding written agreement(s).

4.3.6 Knowledge representation

Knowledge representation shall use ontology authoring frameworks such as the web ontology language (OWL)[10] or similar paradigms, including JSON,[7] XML[8] and RDF.[9] A reasoner can be utilized to ensure logical consistency.

5 Technical and organizational recommendations and requirements for data formats

5.1 General

Data formats can be structured in different ways. These structures depend on the procedure by which the data are generated, the intended use of the data, and the amount of metadata required for interpreting the data properly[27][28] and making it findable, accessible, interoperable and reusable ("F-A-I-R").[2]

5.2 Organizational responsibilities

The organization(s) responsible for establishment, maintenance, and/or potential changes in the ownership of the data format shall be documented. Appropriate contact information (e.g. email addresses, websites) shall be provided for each corresponding organization.

The information regarding the data format shall include at least:

a) format description;

b) version;

c) structure;

d) data representation.

The maintainer of the data format shall take the responsibility for:

— user requests;

— format updates;

— error corrections in the specifications.

Data representation and formatting for all entities, attributes, processes, and features of the same data type shall be consistent throughout the format.

5.3 Documentation

Comprehensive documentation of the following shall be provided on:

a) provenance;

b) maintenance;

c) format structure;

d) data items;

e) data formatting;

f) features of the format.

Any of these properties shall be disclosed appropriately. A stable and identifiable source shall be provided where information on the data format is maintained and updated.

The representation of data types and their metadata shall be documented.

NOTE Documentation can be provided electronically as a file, online or available as a scanned printed document (preferably as a PDF).

5.4 Versioning and change log

A data format shall contain information on the exact format version used (and its subversion(s), if applicable). Stable formats with simple structure (e.g. FASTA), where no changes to the format are to be expected, may be exempt from this rule. In this case an annotation of the exemption shall be included.

Changes made to the format shall be documented and indicated by a change of the version number. Metadata describing the format should be linked to the data it contains.

5.5 Compatibility

Forward compatibility shall be ensured for prior versions of a format. Backward compatibility to prior versions should be provided, if applicable.

5.6 Extensibility

The addition of new data items should be made possible for future versions of a format without affecting the compatibility.

5.7 Compression

For compressed data records, the compression algorithm or an appropriate tool for compression and decompression shall be referenced. For custom compression techniques, the integrity/fidelity of both compression and decompression tools shall be ensured by the data provider or maintainer. In this case, the complete compression/decompression algorithm shall be disclosed. Decompression should maintain the original data integrity.

5.8 Structural and control elements

Any elements with special meanings (e.g. field or record separators, escape sequences, line breaks or similar) shall be documented.

5.9 Requirements for data types within formats

5.9.1 General

Where applicable, the representation of data types should follow commonly accepted standards (IEEE, ISO, etc.).

5.9.2 Encoding of numerical quantity values

The representation of numerical quantity values shall be documented. Standard formats (e.g. IEEE 754) shall be used, if applicable. If numerical quantity values are represented by character strings, the decimal separators, expression for exponents and prefixes shall be specified.

For non-standard representations, the range of permissible values in the respective format shall be indicated. Non-decimal data shall be clearly indicated (and the specification documented accordingly).

Numerical quantity values shall be denoted as measured, inferred or assumed data, if applicable.

Measured data shall include information on:

a) measurement precision;

b) measurement accuracy;

c) measurement uncertainty;

d) method used for obtaining the data, if available.

Applicable quantities/quantitative data shall be assigned to an appropriate unit of measure expressed in SI units, if applicable. If units cannot be expressed in the SI system, appropriate conversion factors shall be provided, if available. Ordinal numerical data shall be indicated as such and appropriate ranges or permissible values shall be specified, if applicable.

5.9.3 Encoding of character strings

For data represented by character strings (legible data) the encoding (e.g. the ISO/IEC 8859 series, Unicode) shall be specified, unless encoding is used in accordance with ISO/IEC 646 (ASCIII).

5.9.4 Encoding of sequence data

Nucleic acid and protein sequence data shall be encoded according to the recommendations of the International Union of Pure and Applied Chemistry (IUPAC) and the International Union of Biochemistry and Molecular Biology (IUBMB) in the “Biochemical Nomenclature and Related Documents” (known as the White Book), released by the IUPAC-IUBMB Joint Commission on Biochemical Nomenclature and Nomenclature Commission of IUBMB.[13]

5.9.5 Time data

The presentation of dates and times shall be in accordance with the formats specified in ISO 8601-1 and ISO 8601-2. Arbitrary time data such as time series data can either be represented in accordance with ISO 8601-1 and ISO 8601-2 or as measured data as specified in 5.9.2.

5.9.6 Boolean data

The representation of Boolean states and permissible values shall be specified. Where applicable, 0 or 1 should be used to represent Boolean states, where 0 means “false” and 1 means “true.”

5.9.7 Biological Imaging data

Images shall be encoded in a recognized standard format, preferably as raw data using a lossless format preserving all imaging data, e.g. TIFF (Tag Image File Format, see ISO 12639). If image size has to be reduced, established image compression formats like JPEG (see ISO/IEC 10918-1) or PNG (see ISO/IEC 15948) can also be used instead. Proprietary image formats (e.g. from microscopy manufacturers) might be used, if convertible into standard formats.

For image data from the health and biomedical domain the standard format for Digital Imaging and Communications in Medicine (DICOM) (see ISO 12052) or similar widely used formats should be considered.

5.10 Consistency and compatibility

Internal coherence and compatibility of data representations, relationships between data elements (and their corresponding metadata), terminologies and the vocabularies used shall be ensured by appropriate means of verification. In addition, structural and semantic interoperability of a format with existing formats in the same domain of life sciences or technologies shall be ensured by appropriate means of verification. Interoperability with formats of technology fields that have a point of intersection or potential points of contact to the given field shall be checked.

Data and semantic context should not be lost during format conversions. However, the loss of accuracy for data and semantic context, should it occur, shall be documented. The internal coherence and integrity of data representations, relationships between data elements (and their corresponding metadata), terminologies and vocabularies shall be ensured by appropriate means of verification. Structural and semantic interoperability of a format with existing formats of the same life science technology field and with formats of technology fields that have a point of intersection or potential points of contact to the given field shall be validated.

The data structure of a format shall be developed in accordance with respective defined minimum reporting guidelines applicable in the corresponding field (see Clause B.2 for examples). The contextual and biological meaning of comprised items, processes and features shall be defined by annotations to domain specific ontologies, taxonomies and controlled vocabularies (See Clause B.3 for examples), using unique PIDs for the corresponding biological and conceptual entities that refer to corresponding resources resolving the entity definition.

5.11 Data integrity

The data format should enable checking of format and data integrity, e.g. by using checksums or other similar methods.

5.12 Format validation

The publisher or maintainer of a format should enable checking of the data format with regards to strength, weakness, applicability, and limitations of the data format, and provide a means of enabling validation of a data file against the format specification.

5.13 Data provenance

The complete history of the data should be documented using structured, interoperable, and hence machine-actionable provenance information. An uninterrupted chain of provenance information should be maintained by linking together metadata describing any preceding processing steps, methods, tools, biological entities, biological material and data utilized to generate the data documented. A complete chain of provenance information enables both, the assessment of data quality and its fitness for a particular purpose and establishing reliability and reproducibility of the data by tracing its origin, generation, processing and analysis. Appropriate precautions, such as access control mechanisms, shall be taken, if the provenance information can contain sensitive or personally identifiable information. A defined model, corresponding serializations and other supporting definitions to enable the inter-

operable interchange of provenance information in heterogeneous environments such as the Web should be established for all data in every format (e.g. the W3C PROV standard[14]).

6 Semantic recommendations and requirements for data formats

6.1 General

Formats shall be clearly structured. They should facilitate extraction of the data and availability of metadata. Data representation and formatting shall be consistent and uniform for all items, processes and features of the same data type. The representation of data types and their metadata shall be documented. Where applicable the representation of data types should follow commonly accepted standards (IEEE, ISO, etc.).

The ability to locate data depends on a well-defined set of biological descriptors. A consistent method of describing a biological entity shall be represented so that the same search, analysis and mining tools can locate the descriptive data across the entire range of life science data domains. Requirements are defined in this standard for developing consensus formats for domain-independent annotation of biological data. These biological descriptors shall be compatible syntactically and semantically across life sciences data domains to ensure data sharing, holding, access, use and reuse.

Entities shall be annotated with the most specific term available in the relevant bio-ontology presented as a URI. If a precise term is not available, then the nearest more general term shall be used. If a sufficiently precise term is not available, then the annotator should request that the more precise term be added by the relevant bio-ontologies maintenance group. In addition, the human readable version(s) of the term should be included as a comment.[15]

If a concept occurs in multiple ontologies, then the ontology most relevant to the data set should be used. For example, tissue types are often defined in multiple ontologies, and the ontology used should be the closest available species-specific ontology.[16][17]

6.2 Minimum consensus information for annotation of biological data

6.2.1 General

The minimum required annotation describes the contextual semantics of a data set, its parameters and results, including the biological, medical and environmental context. This metadata annotation should concisely describe both the basic objective (e.g. problem addressed) of the process that produced the data set (e.g. the analytical or experimental procedure), and its context, components, independent (controlled, varied) quantities and dependent (emergent) measurables. These annotations can be reified as a simple table; converted into a standard semantic format, such as the RDF,[9] or written to be compatible with the "Linked Open Data" concept of the W3C Data Activity.[18] The syntax for the annotations is a series of triplet phrases of the form subject – predicate – object. For example; "liver" – "is a(n)" – "organ". Table 2 lists examples of predicates that should be used for data annotations. The exact syntax and reification shall be "fit for purpose." For example, for data search and retrieval using standard web search engines the syntax shall be suitable for indexing by those search engines.

The required and suggested items for inclusion in the annotation are described below (see also Table 1 for examples of basic required biological descriptors). In addition, the suggested predicate for each item is given. For example, hepatocyte should be annotated with a URI pointing to the corresponding term or entry in a referenced resource, such as controlled vocabulary, domain ontology or terminology, using URIs of resolution services which guarantee the perennial resolvability of the URI where feasible (e.g. the URI referenced in Reference [19] that points and refers to the corresponding term "hepatocyte" in the Foundational Model of Anatomy Ontology (FMA)[20]). Additionally, for being human readable it can be annotated with the respective common names "hepatocyte", "hepatic parenchymal cell", etc.

NOTE Although the RDF framework was used as an example to describe any framework, other frameworks (e.g. JSON[7], XML[8]) can also be used to describe the semantic information.

6.2.2 Species

The annotation shall include a descriptor of the species being studied or processed to be analysed. In cases in which a species designation is not sufficiently precise, then a strain, cultivar, or some other more precise term, should be used, if available. For example, in bacterial studies or microbial biotechnological processes, the bacterial strain, serovar, or molecular subtype should be used in addition to the bacteria name. The predicate used shall be "is". If the annotation cannot exactly identify the correct (sub-)species and therefore refers to a higher-level term in the taxonomy, then the predicate "isVersionOf" shall be used, e.g. "isVersionOf *mammalian*".

6.2.3 Sex

Where applicable, the annotation shall include a descriptor of the sex or gender of the individual being analysed or studied. In cases where sex is not relevant (e.g. in bacteria), this entry shall be "not applicable". If the sex is applicable but unknown, this entry shall be "unknown". The predicate used shall be "is" (see ISO/IEC 5218).

6.2.4 Age

Where applicable, the annotation shall include a descriptor of the age (or the age range) of the individual(s) being analysed or studied. In cases where age is not relevant (e.g. in bacteria), this entry shall be "not applicable". If the age is applicable but unknown, this entry shall be "unknown". The predicate used shall be "is".

6.2.5 Organ

Where applicable, the annotation shall include a descriptor of the organ being analysed or studied. In cases where an organ designation is not applicable (e.g. in bacteria or other microorganisms), this entry shall be "not applicable". If the organ is applicable but unknown, this entry shall be "unknown". The predicate used shall be "is".

6.2.6 Tissue

Where applicable the annotation shall include a descriptor of the tissue being analysed or studied. In cases where a tissue designation is not applicable (e.g. in bacteria or other microorganisms), this entry shall be "not applicable". If the tissue is applicable but unknown, this entry shall be "unknown". The predicate used shall be "is". For complex tissue structures, or if the annotation cannot exactly identify the correct tissue, and, therefore, refers to a higher-level term in a referenced ontology, then the predicate "isPartOf" shall be used, e.g. the "isPartOf *hematopoietic system*".

6.2.7 Cell type

Where applicable the annotation shall include a descriptor of the cell or cells being analysed or studied. In cases where a cell or cells designation is not applicable (e.g. in bacteria or other microorganisms), this entry shall be "not applicable". If the cell types are applicable but unknown, this entry shall be "unknown." The predicate used shall be "is". If the annotation cannot exactly identify the correct cell type and, therefore, refers to a higher-level term in a referenced ontology, then the predicate "isVersionOf" shall be used, e.g. "*leukocyte* isVersionOf *hematopoietic* cell".

6.2.8 Identifiable objects

The annotation shall include descriptors of the relevant identifiable objects in the analytical workflow or study. An identifiable object in an analytical workflow or experiment is any tangible object that can be seen and/or is measured during the process. There can be multiple identifiable objects in an analytical workflow or experiment. The predicate used shall be "is".

6.2.9 Identifiable processes

The annotation shall include descriptors of the identifiable processes in the analytical workflow or experiment producing the data. An identifiable process is any component in the analytical workflow or experiment, which changes over time and is measured; e.g. cell proliferation or death. There can be multiple identifiable processes in an experiment. The predicate used shall be "is". If the annotation cannot exactly identify the correct process and, therefore, refers to a higher-level term in a referenced ontology, then the predicate "isVersionOf" shall be used.

6.2.10 Manipulated entities

The annotation shall include descriptors of the manipulated entities in the analytical workflow or experiment producing the data. A manipulated entity is any component of the analytical workflow or experiment that has been changed and controlled by the experimenter; e.g. by addition of a growth factor to a cell culture or portioning of a population based on a characteristic that differentiates members of the population. There can be multiple manipulated entities in an experiment. The predicate used shall be "is".

6.2.11 Analytical, experimental and computational technology

The annotation shall include a descriptor of the analytical, experimental and/or computational technology used to carry out the annotated process. The predicate used shall be "isVersionOf".

6.2.12 Biological or analytical question

For analytical or experimental processes, the annotation shall include a descriptor of the biological question the analytical workflow or experiment was designed to address. This should describe a high-level biological process such as a disease state, normal homoeostatic control process, developmental process etc. The predicate used shall be "is".

6.2.13 Technology-specific data

The data itself shall be encoded using the relevant domain-specific standard formats for the applied technology (see Annex A for recommended examples). The description of the data shall be encoded in an appropriate domain-specific standard metadata documentation format (see Annex B for recommended examples), following the relevant minimum information standards that provide checklists of metadata objects and their respective metadata attributes to be documented in a certain domain and/or for a certain analytical workflow or experimental setup (see Clause B.2 for recommended examples), and using the appropriate domain-specific ontologies, taxonomies and controlled vocabularies (see Clause B.3 for recommended examples) for specification and annotation of the data.

Examples of basic required biological descriptors are shown in Table 1.

Table 1 — Examples of basic required biological descriptors

Field name (subject)	Predicate	Suggested ontology or vocabulary	Comments	Object – Human readable examples[a]
Species	is	NCBI Taxonomy[21]	The species the experiment was carried out in.	Human, *Escherichia coli*
Sex	is		The sex of the test subject, or of the source tissue or cells, where applicable.	male, female, male–female, female–male, hermaphrodite, and other applicable options
Age	is		Age of the individual the study was done in, or the age of the individual supplying the sample.	2 years, 8 h post fertilization (HPF)

[a] Actual data should also include URI to the specific ontology.

Table 1 *(continued)*

Field name (subject)	Predicate	Suggested ontology or vocabulary	Comments	Object – Human readable examples[a]
Organ	is	FMA[20]	The organ source of the sample.	liver, not applicable
Tissue	is	FMA[20]	The tissue source of the sample.	parenchyma, not applicable
Cell	is	FMA[20], Uberon[22]	The identifiable and/or observable cell types in the sample. Note that the species slot can duplicate this slot in the case of bacteria and other single cell organisms.	hepatocyte
Identifiable Objects	is	Protein[23], GO[24]	Any measured quantities in the experiment. This includes both dependent and independent variables. (dependent variables)	Increase in cell count (dependent), gene deletion
Identifiable Processes	is	GO[24]	What processes can be directly observed in the experiment. (dependent variables)	Cell proliferation, cell death, cell division, small molecule metabolism
Manipulated Entities	is	Protein[23], GO[24], small molecule[25], environment[26]	What experimental qualities are varied in the experiment? (independent variables)	Addition of IL-1, change in nutrient concentration
Experimental Technology	isVersionOf	GO[24], OBI[27]	The technology used in the experiment.	Microarray, cell culture, microscope image
Biological Question	isVersionOf	GO[24], NCI Thesaurus[28]	The basic biological process or biological question the experiment was designed to address.	stimulation of cell proliferation, toxicant effects, embryonic development

[a] Actual data should also include URI to the specific ontology.

Examples of predicates are shown in Table 2.

NOTE Based on the COMBINE/BioModels.net Qualifiers.[29]

Table 2 — Predicate (qualifier) examples

Predicate	Descriptions
is	The biological entity or process represented by the data set element has identity with the subject of the referenced resource. This predicate is used to link the component of the data set to its exact representation in another resource, controlled vocabulary or ontology: e.g. to link a hepatocyte cell in a data set to the term “hepatocyte” in an ontology.
isDescribedBy	The biological entity or process represented by the data set element is described by the subject of the referenced resource. This relation can be used, for instance, to link a species or a parameter to the literature that describes the concentration of that species or the value of that parameter.
hasPart	The biological entity or process represented by the data set element includes the subject of the referenced resource, either physically or logically. For example, this relation can be used to link a cell to the subcellular parts it encloses cell or to link the description of components of a multi-component protein complex.
isPartOf	The biological entity or process represented by the data set element is a physical or logical part of the subject of the referenced resource. This relation can be used to link a data set component to a description of the complex in which it is a part. For example, this relation can be used to link subcellular parts to the enclosing cell or to link the description of a component of a multi-component protein complex to the complex.
isVersionOf	The biological entity or process represented by the data set element is a version or an instance of the subject of the referenced resource. This relation can be used to represent, for example, the 'superclass' or 'parent' of a particular biological entity.

Table 2 *(continued)*

Predicate	Descriptions
hasVersion	The subject of the referenced resource is a version, or an instance of the biological entity or process represented by the data set element. This relation can be used to represent an isoform or modified form of a biological entity, e.g. an isoenzyme of an enzyme class.
encodes	The biological entity or process represented by the data set element encodes, directly or transitively, the subject of the referenced resource. This relation can be used to express, for example, that a specific DNA sequence encodes a particular protein.
hasProperty	The subject of the referenced resource is a property of the biological entity or process represented by the data set element. This relation can be used when a biological entity exhibits a certain enzymatic activity or exerts a specific function.
isEncodedBy	The biological entity or process represented by the data set element is encoded, directly or transitively, by the subject of the referenced resource. This relation can be used to express, for example, that a protein is encoded by a specific DNA sequence.
isHomologTo	The biological entity or process represented by the data set element is homologous to the subject of the referenced resource. This relation can be used to represent biological entities that share a common ancestor.
occursIn	The biological entity or process represented by the data set element is physically limited to a location, which is the subject of the referenced resource. This relation can be used, e.g. to describe a compartmental location, within which a reaction takes place or the organism or organismal part in which a described process takes place.
hasTaxon	The biological entity represented by the data set element is taxonomically restricted, where the restriction is the subject of the referenced resource. This relation can be used to ascribe a species restriction to a biochemical reaction.
isPropertyOf	The biological entity or process represented by the data set element is a property of the referenced resource.

6.3 Syntax and reification

Data may be expressed as a table (as in Table 1), or may reside in a database, or some other data resource. In any case, the annotation should be reifiable to RDF triples[9] as well as to plain text. In all cases, the reification shall be shown to be "fit for purpose".

7 Requirements for terminologies and ontologies suitable for annotation of biological data

7.1 General

Terminologies and ontologies for the description of data, concepts and data entities in the life sciences shall facilitate the identification and understanding of key concepts in the biological or biotechnological domain covered by the ontology.

7.2 Requirements for biological ontologies

7.2.1 Maintainer

The ontology shall have a defined maintainer consistent with the relevant community(ies). The community shall be open to any individual or organization with a vested interest. The organization shall have a web presence.

7.2.2 Maintenance of the ontology

The ontology maintainer shall have a defined set of procedures for maintaining the ontology. This shall include maintenance of the web presence and insurance that the ontology is true to its domain and uses.

There shall be a defined process for adding and removing and/or deprecating terms in the ontology.

7.2.3 Ontology syntax

The maintainer shall define the syntax of the ontology based on the community's needs.

The syntax should be based on an existing ontological infrastructure, such as W3C OWL[10] or other widely used syntax, such as that of the Open Biological and Biomedical Ontology (OBO) Foundry.[30]

7.2.4 Linking to other ontologies and term reuse

The maintainer shall link to other relevant ontologies to the greatest extent possible.

7.2.5 Licensing and attribution

The maintainer shall publish the licensing model for the ontology.

The maintainer shall publish the attribution model for the ontology.

7.2.6 Stable URIs and versioning information

The maintainer shall provide a mechanism to create and share stable URIs for terms and concepts in the ontology.

The maintainer shall provide a mechanism of versioning both individual terms and versions of the entire ontology.

7.2.7 Community involvement

The maintainer shall involve the community most impacted by the ontology in the process of creating, extending and maintaining the ontology.

7.2.8 Language

The human language used in the ontology shall be at the discretion of the maintainer.

The maintainer should make the ontology multilingual to the greatest extent needed and practical. For international use, an English version of controlled vocabularies and ontologies shall be available.

8 Requirements for domain specific data standards

8.1 General

A “domain specific data standard” can refer to either a technology domain defined by the applied methods (microscopy, microarrays, etc.), or a biological domain defined by the underlying processes (cancer, reproduction, etc.).

8.2 Specific requirements for domain specific data standards

8.2.1 Maintainer

The data standard shall have a defined maintainer consistent with the relevant communities. The community shall be open to any individual or organization with a vested interest. The organization shall have a web presence.

8.2.2 Maintenance of the data standard

The data standard maintainer shall have a defined set of procedures for maintaining the standard. This shall include maintenance of the web presence and insurance that the standard is true to its domain and uses.

There shall be defined processes for adding, removing and/or deprecating terms or constructs in the data standard.

8.2.3 Data standard syntax

The maintainer shall define the syntax of the data standard based on the community's needs. The syntax should be based on an existing frameworks such as XML[8], JSON[7] and RDF[9].

8.2.4 Linking to other data standards

The maintainer shall:

a) link to other relevant data standards to the greatest extent possible;

b) reuse data standards from other domains when appropriate;

c) use terms from standards compliant biological ontologies.

The data standard shall be compatible with standards compliant data repositories.

8.2.5 Licensing and attribution

The maintainer shall publish:

a) the licensing model for the data standard;

b) the attribution model for the data standard.

8.2.6 Stable URIs and versioning information

The maintainer shall provide a mechanism:

a) to create and share stable URIs for terms and concepts and for the data standard in total;

b) of versioning the data standard.

8.2.7 Community involvement

The maintainer shall involve the communities most impacted by the data standard in the process of its creation, extension and maintenance.

8.2.8 Language

The human language used in the data standard shall be at the discretion of the maintainer.

The maintainer should try to make the data standard multilingual to the greatest extent needed and practical. For international use, an English version of controlled vocabularies and ontologies should be available.

9 Requirements for data repositories for biological data

9.1 General

Data repositories for biological data shall facilitate the long-term storage, including archiving, indexing, searching and sharing of biological data. Data formats used in data repositories shall be in accordance with this document.

9.2 Requirements for data repositories of biological data

9.2.1 Maintainer

The repository shall have a defined maintainer consistent with the relevant communities and the type of data. The community shall be open to any individual or organization with a vested interest in either creating or using the data. The organization shall have a web presence.

9.2.2 Maintenance of the repository

The repository maintainer shall have a defined set of procedures for maintaining the repository. This shall include maintenance of the web presence and insurance that the repository is true to its domain and uses.

There shall be:

a) a defined process for adding and removing data in the repository;

b) a defined process and schedule for creating backups of the repository.

9.2.3 Repository structure

The maintainer shall define the schema of the repository based on the community's needs.

The data model of the repository should have a consistent schema for all the data throughout the repository and be interoperable with other repositories, e.g. based on an existing F-A-I-R[2] schema infrastructure, if applicable.

The exact schema is at the discretion of the maintainer and the relevant data communities.

9.2.4 Linking to other repositories

The maintainer shall attempt to link to other relevant repositories to the greatest extent possible.

9.2.5 Licensing and attribution

The maintainer shall publish:

a) the licensing model for the repository;

b) the attribution model for the repository.

9.2.6 Stable URIs and versioning information

The maintainer shall provide a mechanism:

a) to create and share stable URIs for data in the repository;

b) for versioning both individual data elements and records, and the entire versions of the repository.

9.2.7 Data visibility

The maintainer shall make the data in the repository visible and accessible to users, to common web search, and indexing engines.

The maintainer should make the data in the repository accessible by human and/or programmatic access or both (e.g. by providing webservices or downloads of the data). Limited access rights can apply.

9.2.8 Community involvement

The maintainer shall involve the communities most impacted by the repository in the process of creating, extending and maintaining the repository.

9.2.9 Language

The human language used in the repository shall be at the discretion of the maintaining organization.

The maintaining organization should attempt to make the repository multilingual to the greatest extent needed and practical. For international use, an English version of controlled vocabularies and ontologies shall be available.

Annex A
(informative)

Examples of common formats for life science data

A.1 General

The formats for the different data types, which are listed in Annex A, are not exclusive. Similar standard formats can also be used if appropriate. Implementations relying on other formats than those listed in Annex A can still be in accordance with this document, if all other requirements and relevant preconditions are fulfilled, i.e. that the applied format is in accordance with the recommendations and requirements of this document. In particular, this applies to data and model formats in rapidly developing technology fields requiring fast adaptation of the concerning data formats.

Online resources are available on the world-wide web summarizing available standards and vocabularies for formatting data in the life-sciences. One widely used example for such a curated, informative and educational resource on data and metadata standards, interrelated to databases and data policies is the publicly available FAIRsharing portal[1)][31][32]. Recommended formats referenced in this document can be found online as a constantly actively curated and updated list in the "ISO 20691 FAIRsharing Collection".[33] More formatting standards and metadata formats than described in this annex can also be found under FAIRsharing and in other online resources.

Formats for medical imaging, medical data recording, electronic patient health records and other person-related health data are mainly out of scope of this document and handled elsewhere. In contrast, formatting and documentation recommendations as described in this document for general data types not solely used in the medical field but also in other fields of the life sciences can also be relevant for the health data domain.

A.2 Data formats for OMICS, biochemical and molecular biology methods

A.2.1 Sequence formats for proteins and nucleic acids

A.2.1.1 General

A number of recognized formats are used for DNA and protein sequences. All formats for sequence data use nomenclatures developed by the International Union of Biochemistry and Molecular Biology (IUBMB) and the International Union for Pure and Applied Chemistry (IUPAC).

Comparison of the amino acid sequences of proteins determined whether relationships between them can have occurred by chance alone.[34] At first done visually for proteins that shared closely related functions, the resolution of gaps and the lack of functional relationships mitigated intuitive rationalization. Visual resolution is superseded by statistical approaches providing algorithms for simple alignment models. Nucleotide sequences were also aligned, analysed and characterized provide models for transcription, translation and codon usage.[35][36]

A.2.1.2 FASTA sequence format

Of the most recognized formats available for sequence data is the FASTA format.[37][38] The FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences, in which nucleotides or amino acids are represented using single-letter codes. The format also allows for

1) This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of this product. Other web resources are also available for summarizing suitable formats and vocabularies.

sequence names and comments to precede the sequences. The FASTA format[39] is simple and lacks facility for extensive annotation. Conversion software from flat files and other formats are readily available and highly conversant avoiding pitfalls associated with ontology.

A.2.1.3 FASTQ sequence and sequence quality format

FASTQ is a text-based file format developed by the Sanger Institute for sharing sequencing data combining both the sequence and an associated per base quality scores in a single file.[40] Platform specific variations of FASTQ representation incorporating scalar differences are interconvertible.

The FAST5 format is based on the hierarchical data format HDF5 format which enables storage of large and complex data.

EXAMPLE The FAST5 format is the standard sequencing output for Oxford Nanopore[2] sequencers such as the MinION.

In contrast to fasta and fastq files, a FAST5 file is binary and cannot be opened with a normal text editor.

Data stored in nanopore FAST5 files can contain the sequence of a read in fastq format (after basecalling) and the raw signal of the pore, as well as several log files and other information.

A.2.1.4 Sequence Read Format (SRF)

The International Nucleotide Sequence Database Collaboration (INSDC) is a foundation initiative supporting a partnership between the DNA Databank of Japan (DDBJ), the European Nucleotide Archive (ENA) of the European Bioinformatics Institute (EMBL-EBI) and the Sequence Read Archive (SRA) of the US National Center for Biotechnology Information (NCBI), NIH's primary archive of high-throughput sequencing data. INSDC provides the DDBJ/ENA/GenBank Feature Table Definition that accommodates EMBL, GenBank and DDBJ sequence formats.[41] Data submitted to any of the three organizations are shared among them.

While gel and capillary based sequencing formats generally require a single trace and recognizably have some metadata included, massively parallel sequence data requires additional analysis and is more functional without metadata. SRA has metadata stripped out. SRA supports input formats from a number of sequencing platforms. There are still developments in the efficiency of conversion for some of the proprietary data formats. SRF has begun to address this variability. SRF is a generic format for DNA sequence data. The primary motivation for creating SRF has been to enable a single format capable of storing data generated by any DNA sequencing technology. Hence, the format has sufficient flexibility to store data from current and future DNA sequencing technologies at minimal cost of implementation.

A.2.1.5 Sequence annotation formats

a) Browser Extensible Data (BED) Format is a flexible format for defining the data lines displayed in an annotation track of the UCSC Genome browser. BED lines have three required fields and nine additional optional fields. The number of fields per line must be consistent throughout any single set of data in an annotation track. The BigBED track format stores track annotations that can be simple or linked to a collection of exons. BigBED files are collections of BED files.[42] This format is under the stewardship of the Global Alliance for Genomics and Health (GA4GH).

b) Wiggle Track Format (WIG) is a line oriented format for graphing in the UCSC browser. WIG has largely been replaced by bigwig.[43]

c) General Feature Format (GFF3)[44] from the Generic Model Organism Database[45] is the most recent and acceptable version of this flat tab-delimited file format addressing shortcomings of previous versions from the Sanger institute. It has nine-tab separate lines. The following situations can be represented in GFF3: canonical genes, non-coding transcripts, parent (part-of) relationships,

2) This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of this product. It is only given as a specific example for the source of the FACT5 format.

alignments, ontology association and database cross references, single exon genes, polycistronic transcripts, genes containing inteins, trans-spliced transcripts, programmed frameshifts and operons.

d) Variant Call Format (VCF) and its binary counterpart binary variant call format (BCF) are text file formats that usually store sequence variants in a compressed form. They can contain meta information lines, a header line and data lines.[46] This format is under the stewardship of the GA4GH. A Variant represents a change in DNA sequence relative to some reference. For example, a variant can represent a Single Nucleotide Polymorphism (SNP) or an insertion. Generally, a single row in a VCF file corresponds to a variant.

e) Gene Transfer Format (GTF) borrows from GFF, but has additional structure warranting a separate definition and format name.[47] Structure is as GFF, so the fields are:

<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes] [comments]

f) Genome Variation Format (GVF) is an extension of Generic Feature Format version 3 (GFF3), is a simple tab-delimited format for DNA variant files, which uses Sequence Ontology to describe genome variation data.[48]

g) Synthetic Biology Open Language (SBOL) is an RDF[9] format for representing, among other things, sequences for genetic circuit designs.[49] It has a rich ability to express both sequence feature annotations and part/sub-part relationships. It is also designed to represent incomplete/partial sequences and relative ordering of parts in a genetic design.

A.2.1.6 Sequence compression formats

CRAM is a compressed columnar file format for storing biological sequences aligned to a reference sequence, initially devised by Markus Hsi-Yang Fritz et al.[50] CRAM was designed to be an efficient reference-based alternative to the Sequence Alignment Map (SAM) (see A.2.2.3) and Binary Alignment Map (BAM) (see A.2.2.4) file formats. It optionally uses a genomic reference to describe differences between the aligned sequence fragments and the reference sequence, reducing storage costs. Each column from the SAM format is separated into its own blocks, improving compression ratio. CRAM files typically vary from 30 % to 60 % smaller than BAM, depending on the data held within them. The CRAM format specification is maintained by the GA4GH.

The ISO/IEC 23092 series (MPEG-G) is a series of ISO/IEC standards providing requirements and recommendations for genomic information representation. The goal of the series is to provide interoperable solutions for data storage, access, and protection across different possible implementations for data information generated by high-throughput sequencing machines and their subsequent processing and analysis. The series utilizes technology and data representation architectures previously validated in the field of digital media. They allow to compress and transport genome sequencing data even in complex scenarios, for instance when access is needed to large amounts of possibly distributed data, or when part of the data needs to be encrypted for privacy reasons. The series is composed of different parts, each one addressing a specific aspect, such as compression, metadata association, application programming interfaces (APIs), and a reference software for data decoding:

a) ISO/IEC 23092-1 specifies data formats for both transport (e.g. streaming) and storage of genomic information, including the conversion process.

b) ISO/IEC 23092-2 provides specifications for the representation of several types of genomic information, such as the syntax and methods for MPEG-G lossless compression of sequencing data and lossy compression of associated quality scores. It only specifies the decoding process and decoder output format while the encoding process is left open to algorithmic and implementation-specific innovations.

c) ISO/IEC 23092-3 specifies metadata storage and interpretation, as well as protection elements providing confidentiality, integrity and privacy rules for the different encapsulation levels as specified in ISO/IEC 23092-1, and defines APIs to access genomic information coded in conformity

to ISO/IEC 23092-1 and ISO/IEC 23092-2. This also includes specifications how to associate auxiliary fields to encoded reads and mechanisms for backward compatibility with existing SAM content (see A.2.2.3), and exportation to this format.

d) ISO/IEC 23092-4 specifies the genomic information representation reference software, referred to as the "genomic model". The decoding software is provided to assess conformity to the requirements of ISO/IEC 23092-1, ISO/IEC 23092-2 and ISO/IEC 23092-6.

e) ISO/IEC 23092-5 specifies a set of test procedures designed to verify whether bitstreams and decoders meet requirements specified in ISO/IEC 23092-1 and ISO/IEC 23092-2. It identifies those requirements, associates them to functionality under test and defines how conformity with them can be tested. Test bitstreams implemented according to those functionalities are provided in electronic form.

f) ISO/IEC 23092-6 provides specifications for the normative representation and coding of annotations for genomic information, such as variants with genotyping information, functional annotations, tracks, expression matrices and contact matrices.

A.2.1.7 Formats for encrypted genomic data

Crypt4GH is a file container format to store genomic data, such as BAM or CRAM, in an encrypted and authenticated state.[51] The approach uses twofold envelope encryption: the data itself is encrypted, and so is the mechanism for unlocking it. The recipient must have their own private key to verify their identity and also a key specific to the file being transferred to access the data therein. The Crypt4GH format is maintained by the GA4GH.

A.2.2 Sequence alignments formats

A.2.2.1 General

Sequence alignment often forms the basis for functional comparison.

A.2.2.2 CLUSTAL-W

CLUSTAL-W Alignment Format is a simple text-based format, often with a *.aln file extension, used for the input and output of DNA or protein sequences into the Clustal suite of multiple alignment programs. Clustal W is well developed and supported by many applications.[52]

A.2.2.3 Sequence Alignment/Map (SAM)

SAM is a TAB-delimited text format consisting of a header section, which is optional, and an alignment section.[53] It is designed specifically for handling large numbers of sequences. If present, the header must be prior to the alignments. Header lines start with "@", while alignment lines do not. Each alignment line has 11 mandatory fields for essential alignment information, such as mapping position, and variable number of optional fields for flexible or aligner specific information.

A.2.2.4 Binary Alignment Map (BAM) format

BAM is the compressed binary version of the Sequence Alignment/Map (SAM) format, a compact and indexable representation of nucleotide sequence alignments.[54] Many next-generation sequencing and analysis tools work with SAM/BAM. For custom track display, the main advantage of indexed BAM over PSL and other human-readable alignment formats is that only the portions of the files needed to display a particular region are transferred to UCSC. This makes it possible to display alignments from files that are so large that the connection to UCSC would time out when attempting to upload the whole file to UCSC. Both the BAM file and its associated index file remain on your web-accessible server (http, https, or ftp), not on the UCSC server. UCSC temporarily caches the accessed portions of the files to speed up interactive display.

A.2.2.5 Multiple Alignment using Fast Fourier Transform (MAFFT)

MAFFT is a high-speed multiple sequence alignment program which implements the Fast Fourier Transform (FFT) to optimize protein alignments based on the physical properties of the amino acids. The program uses progressive alignment and iterative alignment. MAFFT is useful for hard-to-align sequences such as those containing large gaps (e.g. rRNA sequences containing variable loop regions). FASTA and Pearson formats are applicable for MAFFT.

A.2.2.6 Stockholm multiple alignment format

The "Stockholm" format is a system for marking up features in a multiple alignment. These mark-up annotations are preceded by a "magic" label, of which there are four types. The Stockholm format is used by HMMER, Pfam and Belvu.

A.2.2.7 Other sequence alignments formats

Other formats, such as FASTA,[37][38] phyllip[55] and multiple sequence format (MSF), are also used. Interfaces are available to interconvert these formats.

A.2.3 RNA sequence, structure and connectivity formats

Several formats have been used for RNA structure data. For example, the BIOpolymer Markup Language (BIOML), the Bioinformatic Sequence Markup Language (BSMLTM), the Genome Annotation Markup Elements (GAME[56]) and the CORBA Bio effort[57]. These did not have a standard syntax. The RNAML (RNA Markup Language) developed by a consortium of investigators with a reasonably broad representation in RNA bioinformatics is widely used for RNA information files.[58] The ".ct file" contains the nucleic acid sequence and base pairing information from which a structure plot can be computed.

There are currently several variations of the .ct format used for different applications. The formats most widely used are as follows:

— Dot Bracket File Format (DBN): RNA secondary structure is often defined using Dot-Bracket Notation (DBN). Valid structures in DBN format are well-parenthesized words consisting of dots ".", opening "(" and closing ")" parentheses. Dotted positions are unpaired, whereas matching parenthesized positions represent base-pairing nucleotides. As the number of nucleotides interacting is always even (everyone must have a partner), the brackets must be balanced. A structure containing at least two stem-loop structures in which half of one stem is intercalated between the two halves of another stem is called "pseudoknot". The pseudoknot was first recognized in the turnip yellow mosaic virus in 1982.[59] Pseudoknots fold into knot-shaped three-dimensional conformations, but these are not true topological knots, as they are understood in mathematics. Pseudoknots are marked using alternative [..] or { .. b} racket pairs.

— BPSEQ: File name ends with ".bpseq". The bpseq format is a simple text format in which there is one line per base in the molecule, listing the position of the base (leftmost position is 1), the base name (A, C, G, U, or other alphabetical characters), and the position number of the base to which it is paired, with a 0 denoting that the base is unpaired. For more information, see the Comparative RNA Web Site.[60]

  For complexes with more than one molecule, the molecules are listed in sequence, with the base pairs numbers of each successive molecule following, in order, from the previous molecule.

— CT: The first line contains the sequence length L. There are L subsequent lines, one per nucleotide. The i^{th} line starts with i, then the letter denoting the i^{th} nucleotide, then the 5′-connecting base index (i-1), then the 3′ connecting base index (i+1), then the paired base index (or 0 if unpaired), and finally base index in the original sequence. For example, the structure represented in bpseq format above would be represented in ct format as follows:

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

The CT format is applicable for complexes of two or more RNA molecules because the boundaries can be expressed explicitly. If the i^{th} line corresponds to the first nucleotide in one of the molecules, then the third column is 0. If the i^{th} line corresponds to the last nucleotide in one of the molecules, then the fourth column is 0.

This is an example of a short duplex stem, i.e. one composed of two molecules:

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: The syntax allows for the storage and the exchange of information about RNA sequence and secondary and tertiary structures. The syntax permits the description of higher-level information about the data including, but not restricted to, base pairs, base triples, and pseudoknots. A class-oriented approach allows us to represent data common to a given set of RNA molecules, such as a sequence alignment and a consensus secondary structure.

A.2.4 Data formats for mass spectrometry

A.2.4.1 Mass spectrometer output file format (mzML)

mzML is a unified format for LC-MS proteomics data and evolved from mzXML and mzData in the HUPO-PSI working group for mass spectrometry standards.[61] It is well developed and "future proofed". The ProteoWizard software project has number of tools allowing conversion of the XML across a number of platforms including vendor proprietary software deferent to the respective license terms. A stable iteration, mzML 1.1 has been available since 2009.

A.2.4.2 Mass spectrometry-based quantitative studies in proteomics (mzQuantML)

The mzQuantML standard format is intended to store the systematic description of workflows quantifying molecules (principly peptides and proteins) by mass spectrometry. The format was originally developed under the name AnalysisXML as a format for several types of computational analyses performed over mass spectra in the proteomics context, but then has been decided to split into two formats: mzIdentML for peptide and protein identification and mzQuantML. mzIdentML is one of the standards developed by the Proteomics Informatics working group of the Protein Standards Initiative (PSI). mzIdentML does not contain mass spectra. These must be supplied.

TraML is a standard rich XML-format for targeted mass spectrometry method definitions. TraML builds on the same design concepts that were used for mzML and mzIdentML. Like these formats previously developed for different data types, TraML is based on XML and can be parsed and validated for structural correctness with many industry-standard tools.

The proteogenomics formats proBAM and proBed are designed to store a genome-centric representation of proteomics data.

A.2.5 Exchange format for proteomics and metabolomics results (mzTab)

mzTab was developed to bridge the gap between the high level of detail found in XML-based formats required for modelling complete proteomics data and necessary reporting of proteomics and metabolomics data.[62] This format from the HUPO-PSI is ideally suited to make mass-spectroscopy based proteomics and metabolomics results available to a wider biological community outside the field of MS. mzTab addresses the condition of the presence of a specific peptide sequence in a number of different proteins with widely varying structures and functions.

A.2.6 Formats for nuclear magnetic spectroscopy (NMR)

NMR-star is the preferred format for NMR data. It is an extension of the Self-Defining Text Archive and Retrieval (STAR) File, or simply the STAR File. NMR star is a text-based file format for storing structured data. Other formats include PIPP and XEasy.

A.2.7 Format for enzymology data: EnzymeML

EnzymeML is a free and open standard XML-based interchange format for data on enzyme-catalysed reactions. The purpose of EnzymeML is to store and exchange enzyme kinetics data between instruments, software tools, and databases. EnzymeML will allow scientists to share their experimental protocols and results even if they are using different instruments, electronic laboratory notebooks, or databases. EnzymeML is compatible with the Systems Biology Markup Language (SBML). It continues to be evolved and expanded by an international community.

A.2.8 Format for gel electrophoresis: Gel Markup Language (GelML)

Gel electrophoresis is characteristically dependent upon many physical based parameters that are chemically, conformationally and electromagnetically defined in one or two dimensions and differentially.

Metadata are highly significant in gel electrophoresis. GelML, developed within the HUPO-PSI, is a data exchange format for representing gel electrophoresis experiments performed in proteomics investigations.[63][64] Closely following the minimum information about a proteomics experiment (MIAPE), GelML retains several structures from the functional genomics experiment (FuGE) object model.[65]

A.2.9 Formats for real-time PCR data

MIQE (Minimum Information for Publication of Quantitative Real Time PCR Experiments) is a set of guidelines that describe the minimum information necessary for evaluating quantitative real-time PCR experiments (qPCR).[66] This format now also applies to digital PCR.[67]

A.2.10 Formats for genomic sequencing data

The MIxS (Minimum Information About any (X) Sequence) guidelines, developed by the Genomics Standards Consortium (GSC), improves the discoverability of genomic sequence data enabling data integration, discovery and comparison.[68]

MIxS provides core standards for describing genomes, metagenomes and gene marker sequences. It is an extension of the MIGS (Minimum Information About a Genome Sequence) and MIMS (Minimum

Information About a (Meta) Genome Sequence). It currently consists of three separate checklists: MIGS for genomes, MIMS for metagenomes, and MIMARKS for marker genes.

A.2.11 Biomacromolecular structure data

The macromolecular Crystallographic Information File (PDBx/mmCIF) is a dictionary of data archiving macromolecule crystallographic experiments and their results.[69] It is used as official distribution format for formatting and exchanging structural data of macromolecules, such as nucleic acids and proteins. Such structures originally were formatted with one of three crystallographic data formats from the Protein Database (PDB).[70][71][72] A standard tab delimited format was used routinely provided atomic coordinate(s) and bibliographic entry formats (PDB, 1992; PDB, 1996). The PDB Exchange data dictionary (PDBML) provide a schema in XML.[73]

PSI Extended Fasta Format (PEFF) is a unified format for protein and nucleotide sequence databases to be used by sequence search engines and other associated tools (spectra library search tools, sequence alignment software, data repositories, etc.). This format enables consistent extraction, display and processing of information such as protein/nucleotide sequence database entry identifier, description, taxonomy, etc. across software platforms. It also allows the representation of structural annotations such as post-translational modifications, mutations and other processing events. The format has the form of a plain text file that extends the formalism of the individual sequence entries as presented in the FASTA format and that includes a header of meta data to describe relevant information about the database(s) from which the sequences have been obtained (i.e. name, version, etc.).

A.2.12 Small molecule (chemical entities) structures

The IUPAC International Chemical Identifier (InChI) is a textual identifier for chemical substances, designed to provide a standard way to encode molecular information and to facilitate the search for such information in databases and on the web.[74] Initially developed by the International Union of Pure and Applied Chemistry (IUPAC) and the US National Institute of Standards and Technology (NIST), from 2000 to 2005, the format and algorithms are non-proprietary. The InChI identifiers describe chemical substances in terms of layers of information: The atoms and their bond connectivity, tautomeric information, isotope information, stereochemistry, and electronic charge information. Not all layers have to be provided; for instance, the tautomer layer can be omitted if that type of information is not relevant to the particular application. The InChIKey is a hashed version of the full InChI with a fixed length (27 character) condensed digital representation of the InChI that is not human-understandable. The InChIKey specification was released in September 2007 in order to facilitate web searches for chemical compounds. Unlike the InChI, the InChIKey is not unique, but for now there are no known collisions of InChIKeys (no two structures with different InChIs that would have the same InChIKey have been found). The InChI format is used extensively in literature and bioinformatics applications.

Simplified molecular-input line-entry system (SMILES) is an open specification version of the SMILES language, a typographical line notation for specifying chemical structure. It is hosted under the banner of the Blue Obelisk project, with the intent to solicit contributions and comments from the entire computational chemistry community.[75] OpenSMILES is a community sponsored open-standards version of SMILES. SMILES is also widely used in the literature and bioinformatics applications.

An MDL Molfile is a file format for holding information about the atoms, bonds, connectivity and coordinates of a molecule. Each molfile describes a single molecular structure which can contain disjoint fragments. The V3000 molfile and V3000 rxnfile formats are the latest versions of the specification. V3000 is a superset of V2000 in a different format. A connection table (Ctab) contains information describing the structural relationships and properties of a collection of atoms. While still commonly called a molfile, the specification has changed ownership as the companies who first created and then owned the specification changed. MDL Information Systems (MDL) originally developed the format, which is currently owned by Dassault Systemes. Please note that a public, official homepage cannot be found, and therefore an alternative location has been provided.

Structure Data Format (SDF) is a chemical file formats to represent multiple chemical structure records and associated data fields. SDF was developed and published by Molecular Design Limited (MDL) and became the most widely used standard for importing and exporting information on chemicals.

A.2.13 Formats for microarray data and sequencing-based functional genomics

The Functional Genomics Society (FGED) started out with a markup language format for microarray data called MAGE-ML. It was later determined that this product was too complicated for researchers without bioinformatics support. As a result, MAGE-Tab was developed. It is a spreadsheet sheet-based format for microarray gene expression data. The MicroArray Gene Expression Tabular (MAGE-TAB) format is self-contained, does not require an understanding of MAGE-ML or XML. MAGE-TAB can be used for capturing microarray data according to the MIAME guidelines[76] or data from sequencing experiments according to the MINSEQ guidelines.

MIAME (Minimum Information about a Microarray Experiment) is intended to specify all the information necessary for an unambiguous interpretation of a microarray experiment, and potentially to reproduce it. MIAME defines the content but not the format for this information.[77] MIAME/Plant (Minimum Information About a Microarray Experiment involving Plants) is an extension of the MIAME guidelines describing which biological details should be captured for describing microarray experiments involving plants. Detailed information is required about biological aspects such as growth conditions, harvesting time or harvested organ(s).

MINSEQE (Minimal Information about a high throughput SEQuencing Experiment) describes the Minimum Information about a high-throughput nucleotide SEQuencing Experiment that is needed to enable the unambiguous interpretation and facilitate reproduction of the results of the experiment. By analogy to the MIAME guidelines for microarray experiments, adherence to the MINSEQE guidelines will improve integration of multiple experiments across different modalities, thereby maximising the value of high-throughput research.

A.2.14 Formats for glycomics data - Minimum Information Required for a Glycomics Experiment - Mass-spectrometry-based Glycoanalytic Data (MIRAGE MS)

MIRAGE (Minimum Information Required for A Glycomics Experiment) was created to improve the quality of glycomics data in the scientific literature. Researchers seeking to understand the biochemical structure–function relationships of carbohydrates require detailed descriptions of the assay conditions and the experimental results. Currently, these data are insufficiently reported in the literature. A basic description on the sample preparation workflow is required. In contrast to proteomics, different types of glycoconjugates require partially different release approaches, which in turn can have direct influence on the following conditions/parameters: released glycans or still attached to protein/lipid; type of glycan (N-glycan, O-glycan, proteoglycan fragment); and sample pre-treatment prior MS-Analyses (non, reduced, permethylated, endo/exoglycosidase digested, fluorescent label, online/offline LC-separation).

A.2.15 Formats for flow cytometry experiments - Minimum Information about a Flow Cytometry experiment (MIFlowCyt)

The Minimum Information about a Flow Cytometry Experiment (MIFlowCyt) establishes criteria for recording and reporting information about the flow cytometry experiment overview, samples, instrumentation and data analysis. It promotes consistent annotation of clinical, biological and technical issues surrounding a flow cytometry experiment by specifying the requirements for data content and by providing a structured framework for capturing information.

A.2.16 Formats for the description of synthetic biological parts, devices and systems

A.2.16.1 Synthetic Biology Open Language (SBOL)

SBOL is an open standard for the representation of in silico biological designs and their realization through the design-build-test-learn workflow. SBOL Data provides both an electronic format for representing this information (SBOL) and schematic glyphs to graphically depict genetic designs (SBOL Visual).[49]

A.2.16.2 Synthetic Biology Open Language Visual (SBOL Visual)

Synthetic Biology Open Language Visual (SBOL Visual) is an open-source graphical notation that uses schematic “glyphs” to specify genetic parts, devices, modules and systems.

A.2.17 Metadata formats for life science, environmental and biomedical experiments

A.2.17.1 Investigation Study Assay Tabular (ISA-Tab)

ISA-Tab[78][79] describes the ISA (Investigation Study Assay) abstract model reference implementation specified using the ISA-Tab format. ISA-Tab files are tab separated value (tsv) files, with specific labelled column structures.[80] The ISA model consists of three core entities to capture experimental metadata: investigation, study and assay. The extensible, hierarchical structure of this model enables the representation of studies employing one or a combination of technologies, focusing on the description of its experimental metadata (i.e. sample characteristics, technology and measurement types, sample-to-data relationships).

A.2.17.2 Investigation Study Assay JSON (ISA-JSON)

ISA-JSON[81][82] describes the ISA abstract model reference implementation specified using the JSON format,[7] a text format for serializing structured data. The ISA model consists of three core entities to capture experimental metadata: investigation, study and assay. The extensible, hierarchical structure of this model enables the representation of studies employing one or a combination of technologies, focusing on the description of its experimental metadata (i.e. sample characteristics, technology and measurement types, sample-to-data relationships).

A.2.17.3 Fast Healthcare Interoperability Resources (FHIR)

FHIR[83] is designed to enable information exchange to support the provision of healthcare in a wide variety of settings. The specification builds on and adapts practices to enable the provision of integrated healthcare across a wide range of teams and organizations. The intended scope of FHIR is broad, covering human and veterinary clinical care, public health, clinical trials, administration and financial aspects. The standard has been developed by HL7[84] and is intended for global use and in a wide variety of architectures and scenarios.

A.3 Formats for biological imaging

A.3.1 Image Data Resource (IDR)

IDR is a prototype platform for publishing, mining and integrating bioimaging data at scale, following the Euro-BioImaging/ELIXIR imaging strategy using the OMERO and Bio-Formats open source software built by the Open Microscopy Environment.[85] Deployed on an OpenStack cloud running on the EMBL-EBI’s Embassy resource, it includes image data linked to independent studies from genetic, RNAi, chemical, localization and geographic high content screens, super-resolution microscopy and digital pathology.

A.3.2 Open Microscopy Environment eXtensible Markup Language (OME-XML)

The Open Microscopy Environment (OME) develops open-source software and data format standards for the storage and manipulation of biological microscopy data. It is a joint project between universities, research establishments, industry and the software development community. The purpose of OME-XML is to provide a rich, extensible way to save information concerning microscopy experiments and the images acquired therein.

A.3.3 Open Microscopy Environment Ontology (OME-OWL)

The Open Microscopy Environment Ontology (OME-OWL) is a light microscopy imaging ontology that has been developed through translation of the OME data model. The aim of this ontology is to support multi-modal imaging technologies and integrate life-science metadata to enable comprehensive image analyses. The core concepts extracted from the OME Data model include project, experiment, instrument, image, screen, plate and region of interest. The ontology has been extended to include electron microscopy, X-ray computed tomography (CT) and magnetic resonance imaging (MRI).

A.3.4 Open Microscopy Environment - Tag Image File Format (OME-TIFF)

OME-TIFF is a standardized file format for multidimensional image data. OME-TIFF was created to maximize the respective strengths of OME-XML and TIFF. It takes advantage of the rich metadata defined in OME-XML while retaining the pixels in multi-page TIFF format for compatibility with many more applications.

A.4 Formats for computer models of biological systems

A.4.1 CellML

CellML is a machine-readable, XML-based[23] model description and exchange format for computer-based mathematical models.[86][87] CellML is a description language to define models of cellular and subcellular processes. It defines lightweight XML constructs that group mathematical relationships within modules. The variables used in the mathematics are defined within each module, and connections between variables in different modules can be specified. CellML supports component-based modelling, allowing models to import other models, or subparts of models, therefore strongly encouraging their reuse and facilitating a modularized modelling approach. A CellML model typically consists of components, which can contain variables and mathematics that describe the behaviour of each component. The format provides means to reuse and group components into hierarchical structures. All entities (elements) carry an identifier and mathematical definitions are encoded using MathML. The mathematical model is considered to be the primary data, and biological context is provided by annotating the variables and equations with metadata using the RDF.[9][88]

A.4.2 Systems Biology Markup Language (SBML)

SBML is a machine-readable, XML-based[8] model description and exchange format for computational models of biological processes.[89][86] Its strength is in representing phenomena at the scale of biochemical processes, but it is not limited to this only. The evolution of SBML proceeds in stages in which each “Level” is an attempt to achieve a consistent language at a certain level of complexity. Since SBML Level 3 the format is modular, with the core usable in its own right and packages being additional “layers” adding features to the core. By itself, SBML core is suited to representing such things as classical metabolic models and cell signalling models, involving well-mixed substances and spatially homogeneous compartments where they are located. SBML packages that extend the core and are optional in their use, add additional model features, such as visualizations, constraint-based models, hierarchical model composition, or grouping of elements. SBML models are decomposed into explicitly labelled constituent elements. A valid model can consist of various user-defined elements, e.g. substances, products and modifiers involved in processes, or compartments and where these are located.

Biological and structural context of the model, as well as of its content and environment is provided by annotating the model, its parameters, variables, equations, entities, components and other content with metadata using the RDF.[9][88]

A.4.3 Neuroscience eXtensible Markup Language (NeuroML)

NeuroML is a machine-readable, XML-based[8] model description and exchange format for computational models in neuroscience. It was created to facilitate data archiving, data and model exchange, database

creation, and model publication in the neurosciences. The focus is on models which are based on the biophysical and anatomical properties of real neurons.[86][90]

A.4.4 Pharmacometrics Markup Language (PharmML)

PharmML is an exchange format for nonlinear mixed effect models used in pharmacometrics and provides means to encode models, trial designs and modelling steps. PharmML allows for a smooth exchange of computer models between different software tools used in population pharmacokinetics and pharmacodynamics.[86]

A.4.5 Human Physiome Field Markup Language (FieldML)

FieldML[91] is a machine-readable, XML-based[8] model description and exchange format for representing hierarchical models using generalized mathematical fields. FieldML can be used to represent the dynamic 3D geometry and solution fields from computational models of cells, tissues and organs.[86][87]

A.4.6 Biological PAthways eXchange (BioPAX)

BioPAX is a machine-readable standard format that aims to enable integration, exchange, visualization and analysis of biological pathway data.[92]

A.4.7 Systems Biology Graphical Notation (SBGN)

The Systems Biology Graphical Notation (SBGN) is an effort to standardize the graphical notation used in diagrams and maps of biological processes.[86][93] The mission of SBGN is to develop high quality, standard graphical languages for representing biological processes and interactions. The use of a standard visual notation is vital to ensure that diagrams and pathway maps are unambiguous and consistent. Each SBGN language is based on the consensus of the broad international SBGN community of biologists, curators and software developers. SBGN comprises three languages, which allow biological networks to be viewed from different perspectives and at different levels of detail:

— SBGN Process Description language: The SBGN Process Description (PD) language shows the temporal courses of biochemical interactions in a network. It can be used to show all the molecular interactions taking place in a network of biochemical entities, with the same entity appearing multiple times in the same diagram.

— SBGN Entity Relationship language: The SBGN Entity Relationship (ER) language allows to see all the relationships in which a given entity participates, regardless of the temporal aspects. Relationships can be seen as rules describing the influences of entities nodes on other relationships.

— Activity Flow language: The SBGN Activity Flow (AF) language depicts the flow of information between biochemical entities in a network. It omits information about the state transitions of entities and is particularly convenient for representing the effects of perturbations, whether genetic or environmental in nature.

A.5 Formats for model simulations and their results in the life sciences

A.5.1 Simulation Experiment Description Markup Language (SED-ML)

SED-ML is a machine-readable, XML-based[8] exchange format for encoding computational model simulation setups, to ensure exchangeability and reproducibility of simulation experiments.[86][94] It follows the requirements defined in the MIASE guidelines (see Clause B.2).

SED-ML allows the exact simulation set-up to be configured and re-run. It evolves in Levels and Versions and is not specific to any simulation software or modelling format.

A.5.2 Open Modeling EXchange format (OMEX)

The Open Modeling EXchange format (OMEX) supports the exchange of all the information necessary for a modelling and simulation experiment in the life sciences.[95] An OMEX file is a ZIP container that includes a manifest file, an optional metadata file, and the files describing the model. The manifest is an XML[8] file listing all files included in the archive and their type. The metadata file provides additional information about the archive and its content. Although any format can be used, an XML[8] serialization of the RDF[9] is the best option.[95]

A.5.3 Numerical Markup Language (NuML)

NuML is a machine-readable, XML-based[8] format for describing and exchanging multidimensional arrays of numbers to be used with model and simulation descriptions.[86] NuML was initially developed as part of the Systems Biology Results Markup Language (SBRML).[86]

A.6 Descriptors for quality measurements for data and models

Quality Control Markup Language (qcML) is a machine-readable, XML-based[8] exchange format for quality-related data of mass spectrometry that follows the design principles of the related mzML, mzIdentML, mzQuantML, and TraML standards from the HUPO-PSI (Proteomics Standards Initiative). It constitutes a data format geared towards capturing quality control (QC) data from high-throughput experiments and processes in the life sciences. The current focus of the project is towards mass spectrometry-based proteomics, but the format is suitable for metabolomics and next-generation sequencing as well.

Annex B
(informative)

Minimum reporting standards for data, models and metadata

B.1 General

Annex B provides a list of applicable minimum information standards, as well as a list of applicable domain specific ontologies, taxonomies and controlled vocabularies to describe data sets in the life science domains, their contained data elements and their context. The reporting standards listed in Annex B are not exclusive and precluding. Similar standards can also be used, if appropriate.

Implementations relying on other standards rather than the listed ones can still be applied in accordance with this document, providing that all other requirements of this document and relevant preconditions are fulfilled (see Clauses 4 to 8).

Online resources are available on the world-wide web summarizing available standards and vocabularies for describing data in the life-sciences. One widely used example for such a curated, informative and educational resource on data and metadata standards, interrelated to databases and data policies is the publicly available FAIRsharing portal[3)][31][32] Minimum information standards listed in Clause B.2 and terminologies listed in Clause B.3 can be found online as an actively curated and updated list within the "ISO 20691 FAIRsharing Collection".[33] More standard formats and metadata formats can also be found under FAIRsharing and in other online resources.

The list of ontologies in Clause B.3 is an initial collection of artefacts retrieved from the NCBO BioPortal[96] and the EMBL-EBI Ontology Lookup Service[97] deemed to be relevant based on their active use in data management resources for life science data such as the widely used SEEK platform[98] and its FAIRDOMHub installation[99], as well as others.

B.2 Minimum information standards

Table B.1 — List of minimum information standards

Acronym	Name	Description and homepage
CIMR	Core Information for Metabolomics Reporting	Minimal requirements for metabolomics experiments.
CONSORT	Consolidated Standards of Reporting Trials	Reporting of parallel-group randomized controlled trial (RCT), enabling readers to understand a trial's design, conduct, analysis and interpretation, and to assess the validity of its results.
MIABE	Minimum Information About a Bioactive Entity	Reporting requirements for the publication of data on one or a series of bioactive entities, such as pharmaceuticals and pesticides, and their interactions with one or more target molecules.
MIABIS	Minimum Information About BIobank data Sharing	Minimum information required to initiate collaborations between biobanks and to enable the exchange of biological samples and data. The aim is to facilitate the reuse of bio-resources and associated data by harmonizing biobanking and biomedical research.
MIACA	Minimum Information About a Cellular Assay	Information guideline and a modular Cellular Assay Object Model (CA-OM) that can cover the range of cellular assays possible and which is the basis for efficient data exchange.

3) This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of this product. Other web resources are also available for summarizing suitable formats and vocabularies.

Table B.1 *(continued)*

Acronym	Name	Description and homepage
MIAME	Minimum Information About a Microarray Experiment	Information needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment.
MIAPAR	Minimum Information About a Protein Affinity Reagent	Guideline for experimentalists who wish to unambiguously describe protein affinity reagents and their protein targets. It specifies the minimum amount of information required to describe either the production or use of affinity reagents such as antibodies, aptamers, protein scaffolds, etc.
MIAPA	Minimum Information for A Phylogenetic Analysis	Detailed checklist of metadata necessary for researchers to evaluate or reuse a published phylogeny.
MIAPE	Minimum Information About a Proteomics Experiment	Minimum set of information about whole proteomics experiments that would be required by a public repository.
MIAPE-MS	Minimum Information About a Proteomics Experiment - Mass Spectroscopy	Minimum information required to report the use of a mass spectrometer in a proteomics experiment, sufficient to support both the effective (re-)interpretation and (re-)assessment of the data and the potential reproduction of the work that generated them.
MIAPE-MSI	Minimum Information About a Proteomics Experiment - Mass Spectroscopy Informatics	Minimum information required to report the use of protein and peptide identification and characterization software to analyse the data produced by mass spectrometry experiments, sufficient to support both the effective interpretation and assessment of the data and the potential recreation of the work that generated them.
MIAPE-GE	Minimum Information About a Proteomics Experiment - Gel Electrophoresis	Minimum information to report about the use of n-dimensional gel electrophoresis in a proteomics experiment.
MIAPE-GI	Minimum Information About a Proteomics Experiment - Gel Informatics	Minimum information required to report an informatics analysis performed with gel electrophoresis images, in a manner compliant with the aims as laid out in the “MIAPE Principles” document.
MIAPE-CC	Minimum Information About a Proteomics Experiment - Column Chromatography	Minimal set of information to document a column chromatography experiment.
MIAPE-CE	Minimum Information About a Proteomics Experiment - Capillary Electrophoresis	Minimal set of information to document a capillary electrophoresis experiment.
MIAPE-Quant	Minimum Information About a Proteomics Experiment - Mass Spectrometry Quantification	Minimum information required to report the use of quantification techniques in a proteomics experiment, sufficient to support both the effective interpretation and assessment of the data and the potential recreation of the results of the data analysis.
MIAPPE	Minimum Information about Plant Phenotyping Experiment	Reporting guideline for plant phenotyping experiments covering the description of the following aspects of plant phenotyping experiment: study, environment, experimental design, sample management, biosource, treatment and phenotype.
MIARE	Minimum Information About a RNAi Experiment	Set of reporting guidelines that describes the minimum information to be reported about an RNAi experiment to enable the unambiguous interpretation and reproduction of the results.
MIASE	Minimum Information About a Simulation Experiment	Information necessary to enable the execution and reproduction of numerical simulation experiments, derived from a given set of quantitative models.

Table B.1 *(continued)*

Acronym	Name	Description and homepage
MIFlowCyt	Minimum Information about a Flow Cytometry Experiment	Criteria for recording and reporting information about the flow cytometry experiment overview, samples, instrumentation and data analysis.
MICEE	Minimum Information about a Cardiac Electrophysiology Experiment	Reporting standard developed by an international group of leading experimental teams comprising an explicit minimum set of information deemed necessary for reproduction and utilization of published cardiac experimental electrophysiology research.
MIxS - MIGS/MIMS	Minimum Information about a (Meta)Genome Sequence	Conceptual structure for extending the core information that has been traditionally captured by the INSDC (DDBJ/EMBL/GenBank) to describe genomic and metagenomic sequences. The MIMS extension describes key aspects of environmental context.
MINI	Minimum Information about a Neuroscience Investigation	Information required to report the use of electrophysiology in a neuroscience study.
MIRIAM	Minimum Information Required In the Annotation of Models	Set of guidelines for the consistent annotation and curation of computational models in biology. It is suitable for use with any structured format for computational models.
MISFISHIE	Minimum Information Specification for *In Situ* Hybridization and Immunohistochemistry Experiments	Minimum information to be provided when publishing, making public, or exchanging results from visual interpretation-based tissue gene expression localization experiments such as *in situ* hybridization, immunohistochemistry, reporter construct genetic experiments (GFP/green fluorescent protein, β-galactosidase), etc.
MIxS	Minimum Information about any (x) Sequence	Overarching framework of sequence metadata, that includes technology-specific checklists from the previous MIGS and MIMS standards, provides a way of introducing additional checklists such as MIMARKS, and also allows annotation of sample data using environmental packages.
MIAPepAE	Minimum Information About a Peptide Array Experiment	Checklist of data and metadata that accompany a peptide-array experiment.
STRENDA	Standards for Reporting Enzymology Data Guidelines	Minimum information that is needed to correctly describe assay conditions and enzyme activity data in enzymology.
STROBE	Strengthening the Reporting of Observational Studies in Epidemiology	Created by an international, collaborative initiative of epidemiologists, methodologists, statisticians, researchers and journal editors involved in the conduct and dissemination of observational studies, with the common aim of STrengthening the Reporting of OBservational studies in Epidemiology.
STROBE-nut	Strengthening the Reporting of Observational Studies in Epidemiology—Nutritional Epidemiology	Guideline for reporting nutrition epidemiology and dietary assessment research.

B.3 Domain specific ontologies, taxonomies and controlled vocabularies

B.3.1 Medicine, health and disease

Table B.2 — Specific ontologies, taxonomies and controlled vocabularies for medicine, health and disease

Acronym	Name	Description and homepage
ATC	Anatomical Therapeutic Chemical Classification	Classification of active ingredients of drugs according to the organ or system on which they act and their therapeutic, pharmacological and chemical properties. It is controlled by the World Health Organization Collaborating Centre for Drug Statistics Methodology (WHOCC), and was first published in 1976.
CHMO	Chemical Methods Ontology	Data collection in chemical experiments, material analysis and synthesis.
CMO	Clinical Measurement Ontology	Entities for morphological/physiological measurement records from clinical and model organism research.
DCM	DICOM Controlled Terminology	Controlled terms for DICOM.
DINTO	Drug-Drug Interactions Ontology	Formal representation of drug-drug interaction knowledge.
DOID	Disease Ontology	Description of the classification of human diseases organized by aetiology.
DRON	Drug Ontology	Support comparative effectiveness researchers studying claims data.
ICD-10	International Classification of Diseases Version 10	The ICD is the international standard diagnostic classification for all general epidemiological, many health management purposes and clinical use. ICD-10 is the tenth revision of the International Statistical Classification of Diseases and Related Health Problems (ICD), a medical classification list by the World Health Organization (WHO). It contains codes for diseases, signs and symptoms, abnormal findings, complaints, social circumstances, and external causes of injury or diseases.
ICD-11	International Classification of Diseases Version 11	Created to allow the recording, reporting and grouping of conditions and factors that influence health. It contains categories for diseases, health-related conditions, and external causes of illness or death. The purpose of the ICD is to allow the systematic recording, analysis, interpretation and comparison of mortality and morbidity data collected in different countries or areas and at different times. The ICD is used to translate diagnoses of diseases and other health problems into an alphanumeric code, which allows storage, retrieval and analysis of the data. The ICD has become the international standard diagnostic classification for all general epidemiological and many health management purposes.
NCIt	NCI Thesaurus	Reference terminology for many NCI and other systems. It covers vocabulary for clinical care, translational and basic research, and public information and administrative activities.
MedDRA	Medical Dictionary for Regulatory Activities Terminology	Terminology for regulatory authorities in the pharmaceutical industry during the regulatory process, from pre-marketing to post-marketing activities, and for data entry, retrieval, evaluation and presentation.
MESH	Medical Subject Headings	Controlled vocabulary for indexing journal articles and books in the life sciences serving as a thesaurus facilitating search.
NDFRT	National Drug File Reference Terminology	Reference hierarchy to describe physiologic effects (PE) of drugs.

Table B.2 *(continued)*

Acronym	Name	Description and homepage
OGMS	Ontology for General Medical Science	Entities involved in a clinical encounter including very general terms used across medical disciplines. The scope of OGMS is restricted to humans, but many terms can be applied to a variety of organisms. OGMS provides a formal theory of disease that can be further elaborated by specific disease ontologies.
OBIB	Ontology for Biobanking	Ontology built for annotation and modelling of biobank repository and biobanking administration. It is developed based on subset of Ontology for Biomedical Investigations (OBI) using Basic Formal Ontology (BFO) as top ontology and following OBO Foundry principles. The first version of the ontology is merged of two existing biobank related ontologies, OMIABIS and biobank ontology.
OMRSE	Ontology of Medically Related Social Entities	Covers the domain of social entities that are related to health care, such as demographic information (social entities for recording gender (but not sex) and marital status, for example) and the roles of various individuals and organizations (patient, hospital, etc.).
OMIABIS	Ontologized MIABIS	Ontological version of MIABIS (Minimum Information About BIobank Data Sharing).
RxNORM	RxNORM	Normalized names for clinical drugs and links its names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software.
SNOMED-CT	Systematized Nomenclature of Medicine-Clinical Terms	Systematically organized, computer-processable collection of medical terms providing codes, terms, synonyms and definitions used in clinical documentation and reporting. SNOMED-CT is a reference terminology that can be used to cross-map standardized healthcare languages across healthcare disciplines.
SYMP	Symptom Ontology	Disease symptoms, encompassing perceived changes in function, sensations or appearance reported by a patient indicative of a disease.
UMLS	Unified Medical Language System	UMLS integrates and distributes key terminology, classification and coding standards, and associated resources to promote creation of more effective and interoperable biomedical information systems and services, including electronic health records.

B.3.2 Anatomy

Table B.3 — Specific ontologies, taxonomies and controlled vocabularies for anatomy

Acronym	Name	Description and homepage
AEO	Anatomical Entity Ontology	Description of anatomical structures expanding the Common Anatomy Reference Ontology (CARO).
BSPO	Biological Spatial Ontology	Representation of spatial concepts, anatomical axes, gradients, regions, planes, sides and surfaces.
CARO	Common Anatomy Reference Ontology	Facilitate interoperability between existing anatomy ontologies for different species.
FMA	Foundational Model of Anatomy	Domain ontology that represents a coherent body of explicit declarative knowledge about human anatomy. Its ontological framework can be applied and extended to all other species.
RadLex	Radiology Lexicon	Radiology terminology for radiology practice, education and research.

Table B.3 *(continued)*

Acronym	Name	Description and homepage
UBERON	UBER anatomy ONtology	Integrated cross-species anatomy ontology covering animals and bridging multiple species-specific ontologies. It represents a variety of entities classified according to traditional anatomical criteria such as structure, function and developmental lineage. The ontology includes comprehensive relationships to taxon-specific anatomical ontologies, allowing integration of functional, phenotype and expression data.

B.3.3 Biochemistry

Table B.4 — Specific ontologies, taxonomies and controlled vocabularies for biochemistry

Acronym	Name	Description and homepage
CHEBI	Chemical Entities of Biological Interest	Structured classification of molecular entities of biological interest focusing on small chemical compounds.
CHEMINF	Chemical Information Ontology	Ontology for representing chemical information. In particular, it aims to produce an ontology to represent chemical structure and to richly describe chemical properties, whether intrinsic or computed.
EMO	Enzyme Mechanism Ontology	Concepts describing components of an enzyme and its reaction mechanism including the roles the components play in there.

B.3.4 Cells

Table B.5 — Specific ontologies, taxonomies and controlled vocabularies for cells

Acronym	Name	Description and homepage
CL	Cell Ontology	Ontology for the representation of cell types from the prokaryotic, fungal, and eukaryotic organisms. CL merges information contained in species-specific anatomical ontologies as well as referencing other OBO Foundry ontologies such as the Protein Ontology (PR) for uniquely expressed biomarkers and the Gene Ontology (GO) for the biological processes a cell type participates in.
CBO	Cell Behavior Ontology	Terms/entities for describing multi-cell computational models with focus on existential cell behaviours (spatiality, growth, movement, adhesion, death, etc.) and computational models of those behaviours.
CLO	Cell Line Ontology	Terms to standardize and integrate cell line information and to support computer-assisted reasoning.

B.3.5 Genes, proteins and RNA

Table B.6 — Specific ontologies, taxonomies and controlled vocabularies for genes, proteins and RNA

Acronym	Name	Description
BioPAX	Biological Pathways Exchange	Standard language that aims to enable integration, exchange, visualization and analysis of biological pathway data.
GO	Gene Ontology	Structured vocabulary for use by the research community for the annotation of genes, gene products and sequences. The GO defines concepts/classes used to describe gene function and relationships between these concepts.

Table B.6 *(continued)*

Acronym	Name	Description
MOP	Molecular Process Ontology	Concepts for processes at the molecular level.
OMIT	Ontology for MicroRNA Target	Concepts for data exchange standards and common data elements in the microRNA (miR) domain.
PW	Pathway Ontology	Controlled vocabulary for annotating gene products to pathways.
PPIO	Protein-Protein Interaction Ontology	Structured controlled vocabulary for the annotation of experiments concerned with protein–protein interactions.
PRO	Protein Ontology	Terms/entities for the representation of protein-related entities.
RNAO	RNA Ontology	Terms/entities to capture all aspects of RNA, from primary sequence to alignments, secondary and tertiary structure from base pairing and base stacking to sophisticated motifs.
SO	Sequence Ontology	Structured controlled vocabulary for sequence annotation, for the exchange of annotation data and for the description of sequence objects in databases.

B.3.6 Phenotypes

Table B.7 — Specific ontologies, taxonomies and controlled vocabularies for phenotypes

Acronym	Name	Description and homepage
HP	Human Phenotype Ontology	Structured and controlled vocabulary for the phenotypic features encountered in human hereditary and other disease.
MP	Mammalian Phenotype Ontology	Standard terms for annotating mammalian phenotypic data.
NCBITAXON	NCBI Taxonomy	Ontological representation of the NCBI organismal taxonomy. It does not follow a single taxonomic treatise, but rather attempts to incorporate phylogenetic and taxonomic knowledge from a variety of sources, including the published literature, web-based databases, and the advice of sequence submitters and outside taxonomy experts.
OMIM Ontology	Online Mendelian Inheritance in Man Ontology	OMIM is a comprehensive, authoritative compendium of human genes and genetic phenotypes as well as the relationship between them, which is freely available and updated daily. The OMIM ontology contains terms used within the OMIM database.
OMP	Ontology of Microbial Phenotypes	Community ontology for annotating microbial phenotypes, including bacteria, archaea, protists, fungi and viruses.
PATO	Phenotypic Quality Ontology	Ontology of phenotypic qualities, intended for use in a number of applications, primarily phenotype annotation. This ontology can be used in conjunction with other ontologies such as GO or anatomical ontologies to refer to phenotypes.
ORDO	Orphanet Rare Disease Ontology	Structured vocabulary for rare diseases capturing relationships between diseases, genes and other relevant features.

B.3.7 Experiments

Table B.8 — Specific ontologies, taxonomies and controlled vocabularies for experiments

Acronym	Name	Description and Homepage
BAO	BioAssay Ontology	Concepts for chemical biology screening assays and their results including high-throughput screening (HTS) data for the purpose of categorizing assays and data analysis.