BS ISO 20691:2022

BSI Standards Publication

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

Documents
Document Sharing Hub

Full document details: BS ISO 20691:2022

National foreword

This British Standard is the UK implementation of ISO 20691:2022.

The UK participation in its preparation was entrusted to Technical Committee BTI/1, Biotechnologies.

A list of organizations represented on this committee can be obtained on request to its committee manager.

Contractual and legal considerations

This publication has been prepared in good faith, however no representation, warranty, assurance or undertaking (express or implied) is or will be made, and no responsibility or liability is or will be accepted by BSI in relation to the adequacy, accuracy, completeness or reasonableness of this publication. All and any such responsibility and liability is expressly disclaimed to the full extent permitted by the law.

This publication is provided as is, and is to be used at the recipient's own risk.

The recipient is advised to consider seeking professional guidance with respect to its use of this publication.

This publication is not intended to constitute a contract. Users are responsible for its correct application.

© The British Standards Institution 2022
Published by BSI Standards Limited 2022

ISBN 978 0 580 51300 8

ICS 07.080

Compliance with a British Standard cannot confer immunity from legal obligations.

This British Standard was published under the authority of the Standards Policy and Strategy Committee on 30 November 2022.

Amendments/corrigenda issued since publication

Date	Text affected

Documents
Document Sharing Hub

Full document details: BS ISO 20691:2022

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)

Documents
Document Sharing Hub

© ISO 2022

Full document details: BS ISO 20691:2022

COPYRIGHT PROTECTED DOCUMENT

© ISO 2022

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

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

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub

Full document details: BS ISO 20691:2022

Contents

Page

Foreword ... **v**

Introduction ... **vi**

1 Scope ... **1**

2 Normative references ... **1**

3 Terms and definitions ... **1**

4 Recommendations and requirements for the description of entities and concepts in life science data ... **8**
4.1 General ... 8
4.2 Recommended ubiquitous identifier scheme for biological and conceptual entities ... 8
4.2.1 URI provisions ... 8
4.2.2 IRI provisions ... 9
4.2.3 Relationship between URI and IRI ... 10
4.3 Formatting data and contextual descriptive data (metadata) for biological entities and concepts ... 10
4.3.1 General ... 10
4.3.2 Version control ... 10
4.3.3 Arbitrary Limits ... 10
4.3.4 Character sets ... 10
4.3.5 Machine readability ... 10
4.3.6 Knowledge representation ... 11

5 Technical and organizational recommendations and requirements for data formats ... **11**
5.1 General ... 11
5.2 Organizational responsibilities ... 11
5.3 Documentation ... 12
5.4 Versioning and change log ... 12
5.5 Compatibility ... 12
5.6 Extensibility ... 12
5.7 Compression ... 12
5.8 Structural and control elements ... 12
5.9 Requirements for data types within formats ... 13
5.9.1 General ... 13
5.9.2 Encoding of numerical quantity values ... 13
5.9.3 Encoding of character strings ... 13
5.9.4 Encoding of sequence data ... 13
5.9.5 Time data ... 13
5.9.6 Boolean data ... 13
5.9.7 Biological Imaging data ... 14
5.10 Consistency and compatibility ... 14
5.11 Data integrity ... 14
5.12 Format validation ... 14
5.13 Data provenance ... 14

6 Semantic recommendations and requirements for data formats ... **15**
6.1 General ... 15
6.2 Minimum consensus information for annotation of biological data ... 15
6.2.1 General ... 15
6.2.2 Species ... 16
6.2.3 Sex ... 16
6.2.4 Age ... 16
6.2.5 Organ ... 16
6.2.6 Tissue ... 16
6.2.7 Cell type ... 16
6.2.8 Identifiable objects ... 16

AVA Documents
Document Sharing Hub

© ISO 2022 – All rights reserved

Full document details: BS ISO 20691:2022

6.2.9 Identifiable processes ... 17
6.2.10 Manipulated entities ... 17
6.2.11 Analytical, experimental and computational technology ... 17
6.2.12 Biological or analytical question ... 17
6.2.13 Technology-specific data ... 17
6.3 Syntax and reification ... 19

7 Requirements for terminologies and ontologies suitable for annotation of biological data ... 19
7.1 General ... 19
7.2 Requirements for biological ontologies ... 19
7.2.1 Maintainer ... 19
7.2.2 Maintenance of the ontology ... 19
7.2.3 Ontology syntax ... 20
7.2.4 Linking to other ontologies and term reuse ... 20
7.2.5 Licensing and attribution ... 20
7.2.6 Stable URIs and versioning information ... 20
7.2.7 Community involvement ... 20
7.2.8 Language ... 20

8 Requirements for domain specific data standards ... 20
8.1 General ... 20
8.2 Specific requirements for domain specific data standards ... 20
8.2.1 Maintainer ... 20
8.2.2 Maintenance of the data standard ... 21
8.2.3 Data standard syntax ... 21
8.2.4 Linking to other data standards ... 21
8.2.5 Licensing and attribution ... 21
8.2.6 Stable URIs and versioning information ... 21
8.2.7 Community involvement ... 21
8.2.8 Language ... 21

9 Requirements for data repositories for biological data ... 22
9.1 General ... 22
9.2 Requirements for data repositories of biological data ... 22
9.2.1 Maintainer ... 22
9.2.2 Maintenance of the repository ... 22
9.2.3 Repository structure ... 22
9.2.4 Linking to other repositories ... 22
9.2.5 Licensing and attribution ... 22
9.2.6 Stable URIs and versioning information ... 22
9.2.7 Data visibility ... 23
9.2.8 Community involvement ... 23
9.2.9 Language ... 23

Annex A (informative) **Examples of common formats for life science data ... 24**

Annex B (informative) **Minimum reporting standards for data, models and metadata ... 37**

Bibliography ... 47

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022

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.

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022

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.

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022

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.

© ISO 2022 – All rights reserved

AVA Documents Document Sharing Hub
Full document details: BS ISO 20691:2022

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.

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022

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

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022

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.]

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022

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]

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022

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]

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022

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

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022

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

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022

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.

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022

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

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022

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.

© ISO 2022 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022

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.

© ISO 2022 – All rights reserved
Documents
Document Sharing Hub
Full document details: BS ISO 20691:2022