

This is a preview of "ISO 20691:2022". [Click here to purchase the full version from the ANSI store.](#)

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 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](mailto:copyright@iso.org)  
Website: [www.iso.org](http://www.iso.org)

Published in Switzerland

This is a preview of "ISO 20691:2022". [Click here to purchase the full version from the ANSI store.](#)

## Contents

	Page
Foreword.....	v
Introduction.....	vi
<b>1 Scope.....</b>	<b>1</b>
<b>2 Normative references.....</b>	<b>1</b>
<b>3 Terms and definitions.....</b>	<b>1</b>
<b>4 Recommendations and requirements for the description of entities and concepts in life science data.....</b>	<b>8</b>
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
<b>5 Technical and organizational recommendations and requirements for data formats.....</b>	<b>11</b>
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
<b>6 Semantic recommendations and requirements for data formats.....</b>	<b>15</b>
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

This is a preview of "ISO 20691:2022". [Click here to purchase the full version from the ANSI store.](#)

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
<b>7</b>	<b>Requirements for terminologies and ontologies suitable for annotation of biological data.....</b>	<b>19</b>
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
<b>8</b>	<b>Requirements for domain specific data standards.....</b>	<b>20</b>
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
<b>9</b>	<b>Requirements for data repositories for biological data.....</b>	<b>22</b>
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
	<b>Annex A (informative) Examples of common formats for life science data.....</b>	<b>24</b>
	<b>Annex B (informative) Minimum reporting standards for data, models and metadata.....</b>	<b>37</b>
	<b>Bibliography.....</b>	<b>47</b>

This is a preview of "ISO 20691:2022". [Click here to purchase the full version from the ANSI store.](#)

## 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](http://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](http://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](http://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](http://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.<sup>[1]</sup> 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).<sup>[2]</sup> 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.