TECHNICAL REPORT

ISO/TR 3985

First edition
2021-05

Biotechnology — Data publication — Preliminary considerations and concepts

Biotechnologie — Publication de données — Considérations et concepts préliminaires

Reference number
ISO/TR 3985:2021(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 whom 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 https://www.iso.org/directives-and-policies.html).

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

The explosion of life sciences data (big data) has created a need to digitally locate data from diverse biological assays, obtained in a wide range of laboratories, and from a wide range of experimental protocols. To be able to extract value from big data, it is necessary that the data are “findable”, and that the biology measured in the assay is described in a way that it can be located and interpreted. Data producer’s use of a consistent method to describe the biology that their data represents can greatly improve the use of big data. This single, unified description of biological data facilitates locating and extracting value from an abundance of biological data and return increased value to funding organizations.

Many biotech communities have already developed standard data representations specific to their domain[1]. For example, MIAME[2] in the microarray community, OME/OMERO[3] in the imaging and microscopy communities, SBML[4] in the systems biology and reaction kinetics community, and MIABIS in the biobanking domain[5]. What is lacking is a consistent method of describing the represented biological information so that the same search, analysis and mining tools can locate data across the entire range of life science domains. Consensus and guidance are required and provided in this document for the biotech domain-independent annotation of biological data.

The importance of data sharing as an integral part of biological research is recognized in the research community. As a result, a diverse set of stakeholders has developed the FAIR (Findable, Accessible, Interoperable and Reusable) data sharing principles[7]. The intent of FAIR is to act as a guideline for sharing and enhancing the reusability of data holdings. Many life science funding organizations also place increased emphasis on the importance of data sharing. Some require that data sharing plans are included in grant applications and research contracts, i.e. “data must be made as widely and freely available as possible while safeguarding the privacy of participants and protecting confidential and proprietary data[8].” Data sharing is equally critical for various national and international research and biobank networks. Data sharing is known to encourage diversity of analysis and opinion, the testing of alternative hypotheses and enabling of explorations not envisioned by the original investigators, resulting in increased value to the funding organization.

This document lays out concepts, challenges, issues and benefits that are relevant to developing International Standards for data sharing in life science research and provides an overview for specifying standards and best practices that enable data sharing.

Biotechnology — Data publication — Preliminary considerations and concepts

1 Scope

This document reviews best practices that:

a) respect the existing standardization efforts of life sciences research communities;

b) normalize key aspects of data description particularly at the level of the biology being studied (and shared) across the life sciences communities;

c) ensure that data are "findable" and useable by other researchers; and

d) provide guidance and metrics for assessing the applicability of a particular data sharing plan.

This document is applicable to domains in life sciences including biotechnology, genomics (including massively parallel nucleotide sequencing, metagenomics, epigenomics and functional genomics), transcriptomics, translatomics, proteomics, metabolomics, lipidomics, glycomics, enzymology, immunochemistry, life science imaging, synthetic biology, systems biology, systems medicine and related fields.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

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

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

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

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

3.1
big data
bigdata
extensive *datasets* (3.7) — primarily in the *data* (3.2) characteristics of volume, variety, velocity, and/or variability — that require a scalable technology for efficient storage, manipulation, management, and analysis

Note 1 to entry: Big data is commonly used in many ways, for example as the name of the scalable technology used to handle big data extensive data sets.

Note 2 to entry: Big data includes any data that are aggregated into a repository of much larger size than the component data parts. For example, the collection of abstracts of biological publications represents a big data set with more than 20 million entries.

[SOURCE: ISO/IEC 20546:2019, 3.1.2, modified — "bigdata" was given as an alternative term and Note 2 to entry was added.]

3.2

data

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

[SOURCE: ISO/IEC 2382:2015, 2121272, modified — All three notes were removed.]

3.3

data archiver

archiver

individual or organization responsible for the long-term persistence of data and the access to that data

Note 1 to entry: An archiver receives data from a producer and can be funded by the same or different payer.

3.4

data consumer

consumer

user

individual or organization that uses data as a starting point

Note 1 to entry: In the research domain, a data consumer is a scientist or research group.

Note 2 to entry: In the medical domain, a data consumer can be a physician or patient. In some cases, consumer can also be payer.

3.5

data producer

producer

organization or individual that carries out an experiment or measurement, funded by a *payer* (3.11), and producing a data set

Note 1 to entry: In the research domain producer is typically a researcher, in the commercial domain the producer can be a contract laboratory.

3.6

data publication

publication

any of several forms in which data are made available to a wider community

Note 1 to entry: This includes traditional scientific publications in journals as well as the sharing of data via a public repository such as GENBANK. Data publication is typically, though not always, carried out by an entity dedicated to the collection and dissemination of data, e.g. a *data archiver* (3.3).

Note 2 to entry: The “wider community” refers to data consumers, other than the individuals or organization that obtained the data.

3.7

data set

dataset

identifiable collection of data

[SOURCE: ISO 19115-1:2014, 4.3, modified — “dataset” was given as an alternative term and Note 1 to entry was deleted.]

3.8

data sharing

sharing

making data (e.g. numerical, textual, images) available to, and findable by, others

Note 1 to entry: Data are not truly shared, if they cannot be found.

3.9
data sharing plan
formalized description of how a *data producer* (3.5) will accomplish the task of *data sharing* (3.8)

3.10
metadata
meta-data
data that define and describe other data

[SOURCE: ISO/IEC 11179-1:2015, 3.2.16, modified — "meta-data" was added as an alternative term.]

3.11
payer
organization responsible for funding research

Note 1 to entry: This can be a government organization such as a national research institute, a philanthropic organization, a private research organization or, in the medical case a national or private insurance organization.

3.12
proprietary data
data stored in such a way that by design and implementation they are not accessible to everyone

Note 1 to entry: Proprietary data include, but are not limited to, data proprietary to an organization such as a company, or data proprietary to an individual such as health records.

Note 2 to entry: Proprietary data are the opposite of *public data* (3.13).

3.13
public data
data stored in such a way that by design and implementation they are accessible to everyone

Note 1 to entry: Public data are the opposite of *proprietary data* (3.12).

3.14
regionalization
process of expressing a text or data in a particular human language

Note 1 to entry: This includes not only the textual part of the document but also the date formats and varying usages and meanings of commas (,) and periods (.) in numeric formats.

3.15
reification
expression of data or knowledge in a specific language or syntax

Note 1 to entry: Examples include expressing or converting structured data from one format to another, such as from JSON to XML.

Note 2 to entry: Reification also means making a topic represent the subject of another topic map construct in the same topic map according to ISO/IEC 13250-2:2006, 3.11.

3.16
repurposing
practice of using data in a manner other than which it was originally collected

Note 1 to entry: For example, microscope images originally collected for cell counting purposes might be repurposed and used to measure cell morphology.

4 Abbreviated terms

BBSRC	Biotechnology and Biological Sciences Research Council
ChEBI	Chemical Entities of Biological Interest
DNA	Deoxyribonucleic Acid
EOSC	European Open Science Cloud
EU	European Union
FAIR	Findable, Accessible, Interoperable, Reusable
CASRN	Chemical Abstracts Service Registry Number
HTML	Hypertext Markup Language
MIABIS	Minimum Information about Biobank Information Sharing
MIAME	Minimum Information about a Microarray Experiment
NCBI	National Center for Biotechnology Information
NIH	United States Department of Health and Human Services, National Institutes of Health
OME	Open Microscopy Environment
OMERO	Open Microscopy Environment Remote objects
OSPP	Open Science Policy Platform of the European Union
OWL	W3C Web Ontology Language
PDF	Portable Document Format
PID	Persistent Identifier
POD	Plain Old Documentation
RDF	Resource Description Framework
UCSD	University of California San Diego
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
USA	United States of America
SBML	Systems Biology Markup Language
VEGFa	Vascular Endothelial Growth Factor a
XML	Extensible Markup Language

5 Principles

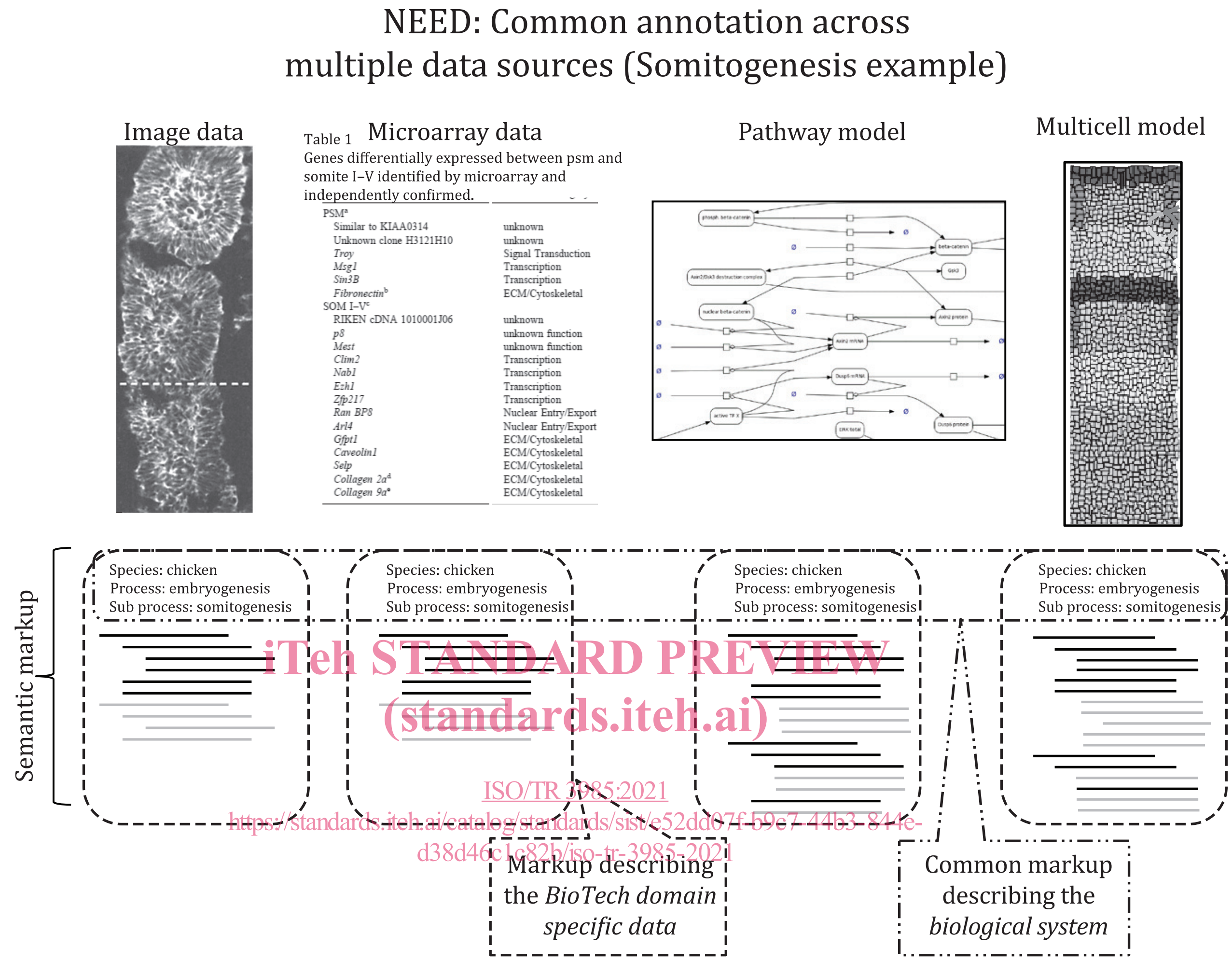
5.1 General

Data sharing by definition is more than simply the publication of summary statistics in tables. It also includes sharing of raw data from which the summaries are generated[8].

The challenge to both researchers and funding agencies is determining what and how data are shared and what metrics might be used to judge the suitability of a sharing plan. For example, the breadth and variety of science supported by the US National Institutes of Health (NIH) prevents the precise content for documentation, its presentation or its transport to be stipulated, i.e. one size does not fit all. As a result, the NIH encourages discussion of data sharing standards and practices between disciplines and professional societies to create a supportive data sharing environment[8].

This view, however, leaves the researcher, reviewer and funding agency without enough guidance and metrics to judge a plan. In addition, it lacks any attempt at standardizing any of the aspects of the data across technology domains, leaving open the potential for ineffective data sharing.

FOUNDATIONAL CONCEPT: At the level of biological description, differences between life science technologies vanish suggesting that a unifying standard spanning all the individual life science data communities can be used for data sharing (See Figure 1).



NEED: Common annotation across multiple data sources (Somitogenesis example)

Image data
Microarray data
Pathway model
Multicell model
Semantic markup
Species: chicken
Process: embryogenesis
Sub process: somitogenesis
Markup describing the *BioTech domain specific data*
Common markup describing the *biological system*

Table 1
Genes differentially expressed between psm and somite I–V identified by microarray and independently confirmed.

PSM[a]	
Similar to KIAA0314	unknown
Unknown clone H3121H10	unknown
Troy	Signal Transduction
Msg1	Transcription
Sin3B	Transcription
Fibronectin[b]	ECM/Cytoskeletal
SOM I–V[c]	
RIKEN cDNA 1010001J06	unknown
p8	unknown function
Mest	unknown function
Clim2	Transcription
Nab1	Transcription
Ezh1	Transcription
Zfp217	Transcription
Ran BP8	Nuclear Entry/Export
Arl4	Nuclear Entry/Export
Gfpt1	ECM/Cytoskeletal
Caveolin1	ECM/Cytoskeletal
Selp	ECM/Cytoskeletal
Collagen 2a[d]	ECM/Cytoskeletal
Collagen 9a[e]	ECM/Cytoskeletal

NOTE In the case shown here four technologies have been applied to the study of somitogenesis, a phase of early embryonic development. Each technology domain (highlighted as - - - -) has its own data and metadata specification. There is a critical need for a common, high level annotation scheme that describes the biology (highlighted as — ·· — ·· —) included in an experiment or model in a (bio)technology-independent fashion.

Figure 1 — Multiple (bio)technologies can be applied to study a biological or biomedical problem.

Consistent annotation of the biological content of data aims at:

a) technology domain independence (i.e. not bound to a certain method or technology);

b) findability of the data;

c) data interoperability (facilitation of data integration);

d) facilitation of data reuse and repurposing.

5.2 Current technologies, approaches and their flaws

Factors that can contribute to the lack of effective sharing and reuse of biological data include:

a) Many communities and their data formats were established before the internet and search engines were available.