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

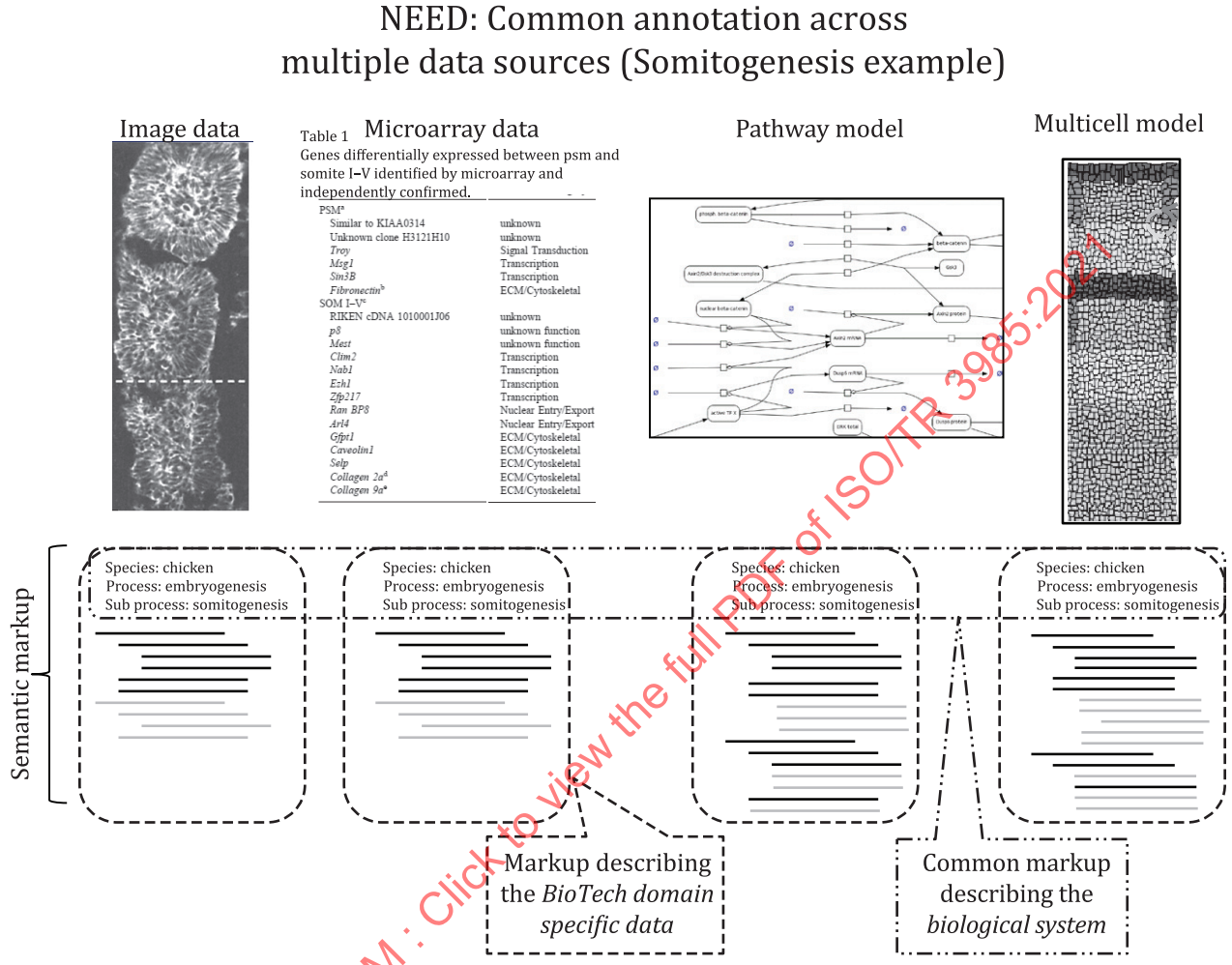
5.1 General

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).



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.

b) The data are not "published" or only partially published (e.g. only available in the form of a summary such as group averages).

c) The data are published in a form, format or location that is not easily interpreted or located.

d) The data are published in a suitable format and at a suitable location, but the terms used in the document are not standard nomenclature making finding the data difficult.

5.3 Standards and best practices to facilitate data sharing and reuse

5.3.1 Maximizing value to the payer

Any standard aiming to facilitate data sharing and reuse maximizes value to the payer by maximizing the number of users and uses of data. Uses of data often extend beyond what was envisioned by the payer and data producer.

5.3.2 Data findability

Data findability by authorized users is essential. In the best case, it will not be necessary to first locate the URI associated with the data source. Current data search applications use artificial intelligence to learn ontological terms that are missing from controlled vocabularies and identify relationships between them, those already in use and query terms to make recommendations for synonyms. Database communities, commercial and public, have already begun to use this type of application. However, data producers can proactively consult ontological databases to determine controlled vocabularies that best render their data findable. Furthermore, "findability" is independent of the biotech domain, and that domain's standard repository, since the life sciences data are described in biological terms and are indexed by web search engines. Searchability of scientific content on the web is covered in Annex A.

5.3.3 Data machine and human interpretability

Life science data can be both machine and human interpretable. The development of "reification technologies" that can present a data document in multiple ways can greatly improve the ability to create documents that are both machine and human interpretable. An example of this technology partially exists in the systems biology domain, especially when it comes to modelling biological processes. Systems Biology Markup Language (SBML)[4] is a free and open interchange format for computer models of biological processes. An SBML format file describes a set of mathematical and chemical equations and can include annotations of the biological data represented in a computational model. The SBML file itself can be executed (a form of reification) by SBML compliant solvers. The SBML document is extremely difficult for a person to manually create, read or understand. However, the SBML can be reified into a human readable document with publication quality formatting, for example into PDF [9]. Native SBML is poorly indexed by web search engines but the PDF reification is easily indexed. This concept can be extended to include other reifications, for example a reification specifically for web publication and ease of indexing by web search engines.

5.3.4 Using accepted controlled vocabularies and naming conventions

It is best practice that data and data descriptions use accepted controlled vocabularies and naming conventions for biological and experimental concepts. The extensive use of existing naming authorities (e.g. gene ontology, CASRN) is critical. In addition, to maintain "backwards compatibility", the inclusion of common synonyms is also best practice.

5.3.5 Biological annotation technology domain independence

When an effort is made to harmonize existing technology domain-dependent data through standardization, the highest-level biological annotation will become technology domain independent.

5.3.6 Data locatability using multiple queries

It is best practice that data are locatable using biological or technology domain-based queries, or both. In the first case, a query such as “hepatocarcinoma” is effective, if it retrieves data across biotechnology domains. In the second case, a query with “microarray” is effective, if it retrieves data across biological domains. Combining the two queries can retrieve the intersection of the biological and technological domains.

5.4 Additional desirable attributes

5.4.1 Data linkage to a published and openly accessible document describing the experimental system

Data linkage to a published and openly accessible document describing the experimental system is important. Without a linkage, the publication of the data in an electronic form can be inhibited. However, in some situations, useable data are not described in a publication, or a publication exists but because of the nature of the data the data themselves are only briefly mentioned in the publication. This situation often occurs when large amounts of data are generated by a project, particularly in ongoing projects that span over many years. In cases like these, the scientific publication has relatively little of the biological details since those details are too extensive for inclusion in a paper. Furthermore, since the data set increases over time, some of the data can be obtained after publication of a scientific paper.

5.4.2 Data format linkage to a published and openly accessible document describing the format

Data are useful only if accessible and/or stored in well structured, consistent formats. This can be achieved by linking the data to a published and openly accessible document describing the format. Individual technology communities are responsible for defining their own standards for their specific data formats.

EXAMPLE It is up to the microscopy community to define the data format standard for microscope images.

5.4.3 Existing information technology

Commercial applications, that ensure the findability and accessibility of massive amounts of data via web services, are already available. Biological data producers can follow the same criteria in making their life science data accessible. Many life science organizations use massive relational databases to store their product information and have developed effective technologies for making the data in those databases indexable and searchable by web search engines. Any large life science data repositories can be handled in the same way. Many countries and regions of the world require sharing and publication of research results, e.g. the initiatives to develop OpenScience or OpenData in the EU[10],[11] and the National Institute of Health's Public Access Policy[12] in the USA. However, these efforts have not yet included standard formats or even the use of controlled vocabularies.

5.4.4 Development of tools and best practices for creating web friendly and search engine crawlable data documents

For big data life sciences, frameworks for annotating and publishing data can support effective data sharing, and value return to payers.

5.5 Essential considerations

5.5.1 Common annotation across multiple data sources

Annotations can minimally include basic concepts, chosen from existing bio-ontologies and controlled vocabularies, presented as keywords. More advanced annotation schemes can embed detailed knowledge such as the relationship between entities and processes in the experiment.

A basic low-level specification is best, if it can include the following parts:

a) terms describing who did the experiment and when;

b) terms linking to any relevant publications;

c) terms describing the overarching biological paradigm of the experiment or data;

NOTE This can include information regarding the “big picture” or driving biological questions for carrying out the experiment.

d) terms describing e.g. the species, sex, age of the test population;

e) biological terms describing:

1) identifiable biological objects such as tissue, cell types, proteins and genes discernible and/or measurable in the experiment;

2) identifiable biological processes such as mitosis or necrosis, measurable in the experiment.

In some cases, the items listed above are not applicable.

In the best cases, the terms above are chosen from standard biological naming authorities or ontologies or both. Some life sciences domain's data standards include the items listed above, but often use different terminology.

Some possible approaches to data annotation and sharing are presented in 5.5.2 to 5.5.4.

5.5.2 Keyword template

An example of simple keyword annotation, using a defined set of keyword “slots”, is shown in Table 1. This example describes the microarray experiment shown in Figure 1 and excludes common annotations such as the person/entity that performed the experiment and a literature citation. Annotations can be included as plain text or as comments in a variety of data file formats.

Table 1 — Simple keyword annotation

Keyword slot name	Human readable	Controlled vocabulary URI/URL
Species:	“*Gallus gallus*”, chicken	http://purl.obolibrary.org/obo/NCBITaxon_9031
Gender:	not applicable	
Life stage or age:	embryo	http://purl.org/sig/ont/fma/fma296970
Organ:	“vertebral column”, spine	http://purl.org/sig/ont/fma/fma13478
Tissue:	not applicable	
Cell:	somite, “presomitic mesoderm”	http://purl.org/sig/ont/fma/fma85522 http://purl.org/sig/ont/fma/fma69072
Biological question:	embryogenesis, somitogenesis	http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#C16649 http://purl.obolibrary.org/obo/GO_0001756
Identifiable objects:	“vascular endothelial growth factor A”, VEGFa	http://purl.obolibrary.org/obo/PR_000017284
NOTE In the "Human readable" column, quotation marks indicate alternate names and commas delimit lists of distinct elements, e.g. one-to-many relationships.		

Table 1 *(continued)*

Keyword slot name	Human readable	Controlled vocabulary URI/URL
Identifiable processes:	"cell division", "cell maturation", "cell differentiation"	http://purl.obolibrary.org/obo/GO_0051301 http://purl.obolibrary.org/obo/GO_0048469 http://purl.bioontology.org/ontology/CSP/0600-1092
Experimental technology:	"gene expression", microarray	http://purl.obolibrary.org/obo/GO_0010467 http://purl.obolibrary.org/obo/OBI_0400147
NOTE In the "Human readable" column, quotation marks indicate alternate names and commas delimit lists of distinct elements, e.g. one-to-many relationships.		

For example, a web accessible document that embedded the above in a search engine friendly format (HTML, DOC, PDF, XLS, plain text and many other formats but not XML) would be "findable" with a web search engine query using the human readable keywords or the controlled vocabulary links (e.g. NCBITaxon_9031 or GO_0048469). Many journals already include a table of abbreviations with each article. Such tables can easily be augmented to include not only the abbreviations used in the paper, but also the proper names of entities (and URI links) as shown in Table 1. A combined table can serve as both the abbreviation list and provide the mappings between the terminology used in the article and standard naming authorities (see Annex B).

5.5.3 Embedding ontological descriptions

Much more complex annotation schemes can be envisioned, that provide information and knowledge beyond the capability of simple keyword annotations, e.g. embedding ontological snippets within data documents. The snippets would be extractable by reification engines. One can imagine modifying existing annotation tools (such as POD or Doxygen, which are used to annotate computer code) so that, besides being able to incorporate human readable (and search engine indexable) annotations, ontological descriptions can also be included. An example for this is the incorporation of a small piece of OWL (the web ontology language) description in Manchester Syntax[13] (a more human readable format than the standard RDF or XML of OWL):

```
## DBBRA: Distributed Biodata and Biomodel Resource Annotations
## DBBRA: OWL Snippet
## DBBRA:  Class:hepatocyte (isA CBO:cell)
## DBBRA:    hasProcess
## DBBRA:     GO:"omega-hydroxylase P450 pathway"
## DBBRA:       isDefinedBy
## DBBRA:         urn:miriam:obo:GO:0097267
## DBBRA: End OWL Snippet
```

The snippet above not only links biological objects to naming authorities (e.g. Gene Ontology), but also defines a knowledge structure stating that the "omega-hydroxylase P450 pathway" is a process that exists in "hepatocyte(s)" which are "cell(s)"[14].

5.5.4 Pseudo-documents

A "pseudo-document" is a data file created extemporaneously in response to a query. Typically, these documents are created from data in one or more relational databases and the document does not exist until a user requests it. Nonetheless, these "non-existent" documents are indexable by web search engines, through designations with unique URIs. The eCommerce community provides massive storage capacity for accessing multiple data files to prepare pseudo-documents ensuring that their stakeholder product lists, maintained in disparate large relational databases, are "findable" by consumers.

PubMed®[1] also uses this approach to generate the documents indexed by web search engines and for the dynamic creation of documents in response to user queries. Several scientific data (and model) repositories use similar approaches. For example, the BioModels database[15] is fully indexed by at least one commonly available large scale internet search engine. A more complex example is the Cell Centered (image) Database[16] and the related Cell Image Library[17].

A search using a commonly available large-scale internet search engine with the query “Alzheimer's serial section 2d image Neocortex pyramidal cell JEOL100CX” locates a “pseudo-page” in the UCSD database containing six relevant images as the first search engine hit.

NOTE Web resources for searching and systems enabling findability are not guaranteed to be persistent and can change over time.

6 Major challenges

6.1 General

The following subclauses list (in no particular order) the current challenges.

6.2 Domain

Challenge: This document provides descriptions of biomedical (human) relevant data publication and sharing. However, the scope of this document covers life sciences e.g. including non-human relevant research such as commercially important animals, plants, etc.

6.3 Regionalization

Challenge: Multiple human languages and character sets pose a challenge, if no commonly understandable language version (e.g. English) is available. Although it is important to have common terminology, the use of multiple languages facilitates harmonization. Many biomedical data publications (and databases) use English, but there are some that do not. In general, ontologies can be regionalized with alternate languages, however this is not generally reflected in practice.

6.4 Proprietary data

Challenge 1: Sharing (e.g. publication, search) of proprietary data within an organization can pose a challenge.

Challenge 2: Sharing (e.g. publication, search) of data and documents that reside behind a paywall, e.g. on a journal publisher's website, can pose an issue for their broader application. If the document is indexed and findable, but access to the full publication requires a subscription (or a single copy fee), it can inhibit its use.

6.5 Large number of existing bio-ontologies, controlled vocabularies and terminologies

Challenge: The unknown extent to which biotechnology communities worldwide are developing or have developed bio-ontologies and domain-specific controlled vocabularies and terminologies is a challenge as it is not openly voiced which new standards are needed.

Comprehensive resources for community standards in the life science domain often list hundreds of such ontologies and terminologies, for example almost 800 are referenced in https://fairsharing.org[18].

1) PubMed® is an example for a free full-text archive of biomedical and life sciences journal literature. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of this product.

6.6 Large number of existing data repositories and corresponding domain specific data formats

Challenge: The unknown extent to which many existing data repositories worldwide are developing or have developed technology-domain specific formats is a challenge as it is not openly voiced which new standards are needed.

Comprehensive resources for community standards in the life science domain often list hundreds of such data formats, for example more than 440 are referenced in https://fairsharing.org[18].

6.7 Large number of funding agencies (e.g. national, educational, philanthropic, commercial)

Challenge: The unknown extent to which funding agencies worldwide, many of which have not developed their own formats, want to be involved in the development of new standards is a challenge as their wish to engage in standardization is not openly voiced.

7 Examples of existing national and regional standards or requirements for data sharing or publication

7.1 General

A comprehensive list of data sharing policies can be found at web resources such as https://fairsharing.org/policies/[18].

7.2 USA

The USA has several research funding agencies. An example of standards and requirements for data sharing and publication is that of the National Institute of Health (NIH). The NIH requires data sharing and publication for NIH funded projects; guidance is provided in Reference [8].

7.3 Canada

Canada has several research funding agencies. An example of standards and requirements for data sharing and publication is the Tri-Agency Open Access Policy on Publications Policy Summary[19]. This policy requires that research papers funded by the agency are freely accessible online and that bioinformatic data (e.g. DNA sequence data) are deposited in the appropriate public database.

7.4 European Union

The Open Science Policy Platform (OSPP) of the European Union (EU) outlines the EU open sharing policy for research data [10]. A study of the importance of effective data sharing in the EU estimated a cost of at least 10,2 billion euros per year for ineffective data sharing [6]. The final report of the EU Open Science Policy Platform on "Progress on Open Science: Towards a Shared Research Knowledge System"[11] provides a brief overview of its mandate, followed by an update on progress by each stakeholder group across the European Commission's eight ambitions on Open Science (OSPP-Recommendations). This summary of practical commitments for implementation with specific examples of progress by each stakeholder community across Europe is followed by a perspective from each group on the major outstanding blockers to progress and possible next steps. The document also proposes a vision for moving beyond Open Science to create a shared research knowledge system by 2030.

The European Open Science Cloud (EOSC) is a European Commission initiative aiming at developing an infrastructure providing its users with services promoting open science practices. EOSC constitutes a 'federated ecosystem of research data infrastructures' that allows the scientific community to share and process publicly funded research results and data across borders and scientific domains. The EOSC FAIR working group published "Six Recommendations for implementation of FAIR practice"[20] and "A Persistent Identifier (PID) policy for the European Open Science Cloud (EOSC)"[21].

7.5 Germany

The Alliance of Science Organisations in Germany supports the long-term archiving of research data, open access to it and compliance with the conventions of individual disciplines in the "Principles for the Handling of Research Data", adopted in 2010. The "Guidelines on the Handling of Research Data" put the framework stipulated by the "Principles for the Handling of Research Data" into a concrete form in the DFG's funding arrangements[22].

7.6 China

In March 2018, the Chinese state council released regulation on scientific data management [23]. These regulations specify that scientific data producers are required to collect scientific data and build a database based on relevant standards. The regulations also indicate that authorities are responsible for establishing rules and regulations while implementing national scientific data policies. Furthermore, the regulations state that, wherever possible, data produced through government funding need to be open to the public.

7.7 United Kingdom

Multiple organizations in the UK have data sharing standards including the medical research council[24], BBSRC[1], Cancer Research UK[25] and Wellcome[26]. In general, these data sharing policies recognize that different fields of study will require different approaches. It does not prescribe when or how researchers are required to preserve and share data.

7.8 India

The India National Data Sharing and Accessibility Policy (NDSAP) is applicable to all shareable data available either in digital or analogue forms generated using public funds[27]. The policies stated objective is to "facilitate access to Government of India owned shareable data through a wide area network, thereby permitting a wider accessibility and usage by public".

7.9 Japan

To promote sharing and utilization of human data, while considering protection of personal information, the National Bioscience Database Center (NBDC) of the Japan Science and Technology Agency (JST) established a platform for sharing various human-related data, the NBDC Human Database, and developed rules and guidelines for operating it[28]. The sharing of policies is summarized as an attempt to maximize:

— collecting human data generated using public funds as much as possible;

— promoting wide sharing of collected data;

— respecting, to the greatest extent possible, the rights of individuals who are research subjects.

In addition, the Japan Agency for Medical Research and Development (AMED) has developed a "Data Sharing Policy for the Realization of Genomic Medicine"[29].

8 Existing legal requirements for data protection

8.1 USA

Health Insurance Portability and Accountability Act (HIPAA).

8.2 European Union

Any organization that processes the personal data of people in the European Union is required to comply with the EU General Data Protection Regulation (GDPR)[30]. Under certain conditions, the GDPR also applies to companies that are not located in Europe.

9 Timing of data publication

Timing of data publication is often defined by the payer and outlined in Clause 7.

10 Costs of data publication

The financial cost of data publication is often a significant concern. Many scientific journals charge a fee for "open access" publication where the document is accessible without a subscription to the journal. For non-standard publication modalities, such as deposition into a public database, there can also be a charge, particularly if the data producer is also the data archiver. Funding agencies (payers) typically recognize these costs and allow for them in a manner that is consistent with their publication and data sharing requirements.

11 Archival data

Archival data can fit into a standard. For example, the 20+ million abstracts already present in PubMed®.

12 Validation and verification of compliance

Validation and verification of compliance are procedures that can be used post-publication that show that a data publication (e.g. scientific journal article, deposit into web-accessible database) conforms to a specified standard.

EXAMPLE Web searches with suitable queries are able to find a specified data document and data can be locatable without needing to know where the data resides. Where access is restricted to only those users with authorized access, permission/s can be required depending on the specific type of data that is accessed.

13 Affected stakeholder categories

A summary of affected stakeholder categories is provided in Table 2.

Table 2 — Summary of affected stakeholder categories

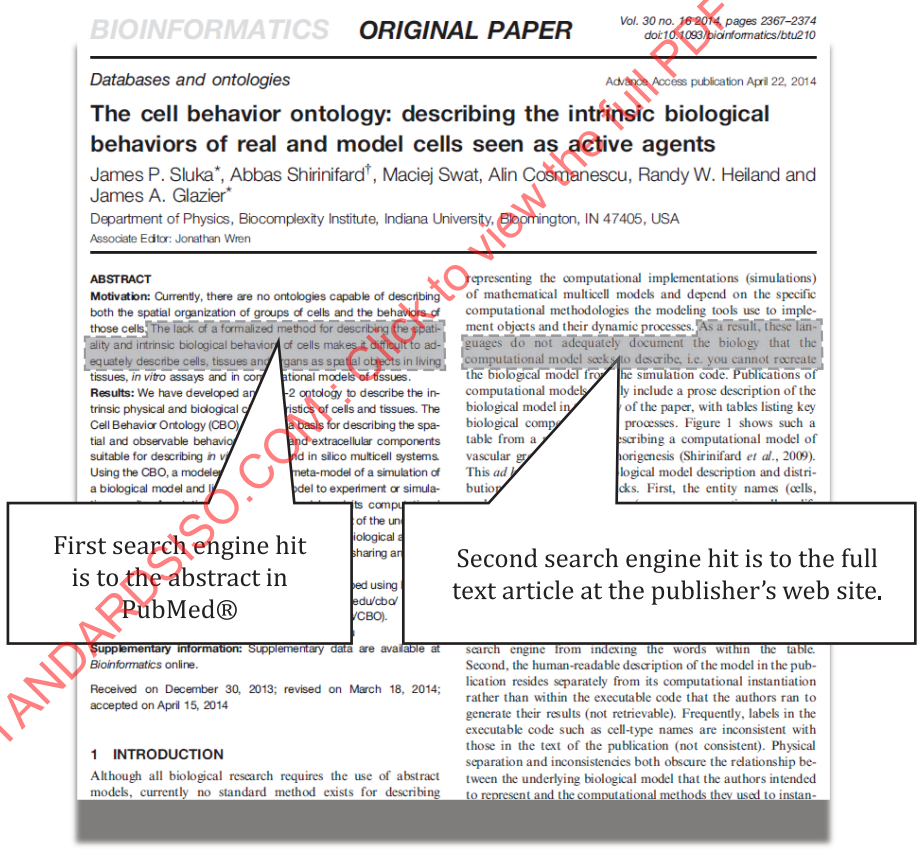
Stakeholder Categories	Benefits or impacts
Industry and commerce	Standardized data formats can be required for data publication.
Government	Data publication standards will significantly increase the value returned for government funded research projects.
Consumers	Increased ability to find relevant biological data increases the consumer's ability to make informed medical and life choices.
Labour	None expected.
Academic and research bodies	Data publication standards can be required by funding organizations.
Standards application businesses	Unknown.
Non-governmental organizations	If they fund biomedical research, then see the benefits or impacts listed for government. If they carry out research using government funding, then see the benefits or impacts listed for academic.

Annex A
(informative)

Searchability of scientific content on the web

A.1 Word and phrase search

A few simple tests were performed to explore how well commonly available large-scale internet search engines index scientific documents. In the first test, shown in Figure A.1, a search with an entire sentence from an article's abstract returns the article's record in PubMed®[2] (www.ncbi.nlm.nih.gov/pubmed). PubMed® does not include the full text of articles, only the title, abstract and bibliographic information. In the second test, also shown in Figure A.1, a search with a commonly available large-scale internet search engine with a sentence from the body of an article returns the full text version of that article (a PDF file) from the publisher's website. This requires that the article is published as open-access and the journal does not have it hidden behind a paywall (therefore making it a part of the "deep web" of sites not indexed/indexable by search engines).



NOTE A search with a commonly available large-scale internet search engine with the highlighted texts locates the article in different repositories. The background image is taken from Reference [31].

Figure A.1 — Scientific content indexed by commonly available large-scale internet search engines

2) PubMed® is an example for a free full-text archive of biomedical and life sciences journal literature. This information is given for the convenience of users of this document and does not constitute an endorsement by ISO of this product.