

Standards for Data acquisition and management

Regulatory & data reporting thematic area

Martin Golebiewski, HITS
Heidelberg Institute for
Theoretical Studies, Germany



Workshop
Putting Science into
Standard "Organ-on-chip:
Toward standardization"

28th-29th April 2021



How to Share Data FAIR ?



Image: Australian National Data Service [ANDS] (<https://www.ands.org.au>) (licensed under a Creative Commons Attribution 4.0 International License)

SCIENTIFIC DATA

OPEN Comment: The FAIR Guiding Principles for scientific data management and stewardship

SUBJECT CATEGORIES
 + Research data
 + Publication characteristics

Box 2 | The FAIR Guiding Principles

To be Findable:

- F1. (meta)data are assigned a globally unique and persistent identifier
- F2. data are described with rich metadata (defined by R1 below)
- F3. metadata clearly and explicitly include the identifier of the data it describes
- F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:

- A1. (meta)data are retrievable by their identifier using a standardized communications protocol
 - A1.1 the protocol is open, free, and universally implementable
 - A1.2 the protocol allows for an authentication and authorization procedure, where necessary
- A2. metadata are accessible, even when the data are no longer available

To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles
- I3. (meta)data include qualified references to other (meta)data

To be Reusable:

- R1. (meta)data are richly described with a plurality of accurate and relevant attributes
 - R1.1. (meta)data are released with a clear and accessible data usage license
 - R1.2. (meta)data are associated with detailed provenance
 - R1.3. (meta)data meet domain-relevant community standards

SCIENTIFIC DATA, 3: 160018 (2016)

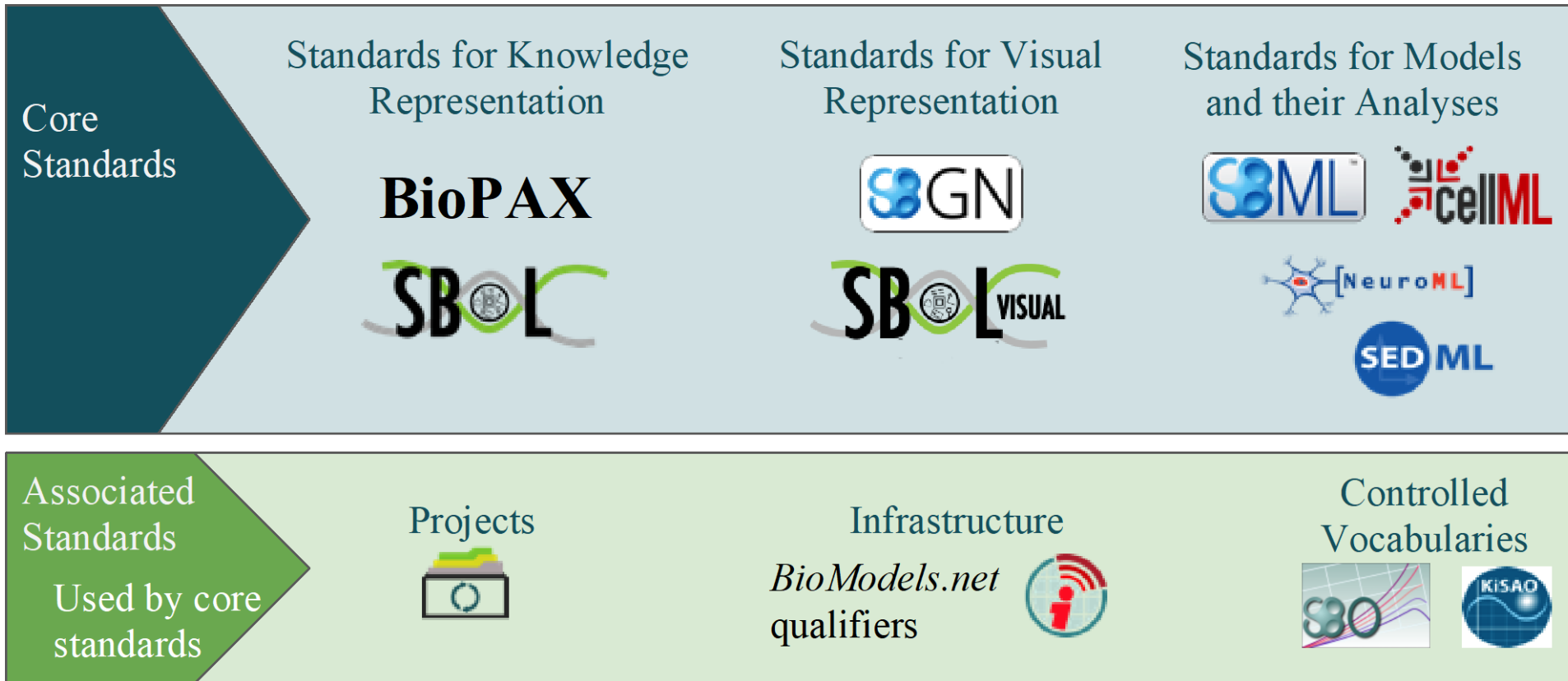
Wilkinson M, Dumontier M, Aalbersberg I, et al.

<https://doi.org/10.1038/sdata.2016.18>

COMBINE Community Standards for Computational Modelling in Biology



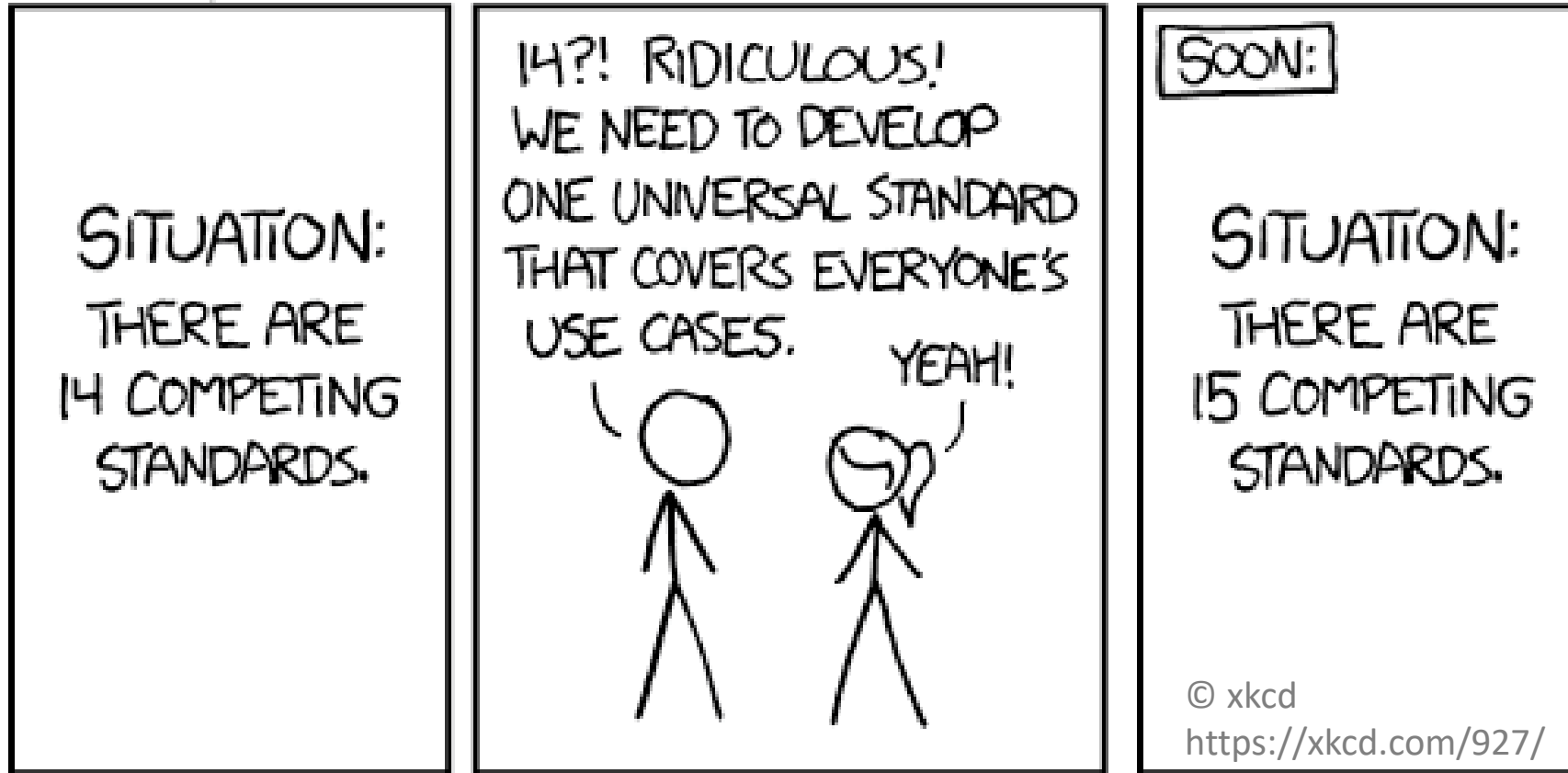
<http://co.mbine.org>



adapted from:
Schreiber F, Bader GD, Gleeson P, Golebiewski M, Hucka M, Le Novère N, Myers C, Nickerson D, Sommer B, Walthemath D: **Specifications of Standards in Systems and Synthetic Biology: Status and Developments in 2016** J Integr Bioinform. (2016) 13:289. doi: 10.2390/biecoll-jib-2016-289

... so many standards

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC.)



Standardized and Harmonized Data Sharing: ISO 20691 (Draft)

Requirements for data formatting and description in the life sciences



Meta-standard for data formatting, description, reporting, integration and sharing

Catalogue of criteria and requirements for life science data formats and semantic data description as prerequisites for a framework of interoperable standards



Example: Great Baltimore fire of 1904

Individual fire hydrants depending on region with 600 variations of hose couplings

Need for a set of harmonized and interoperable data standards

Standardized and Harmonized Data Sharing: ISO 20691 (Draft)

Requirements for data formatting and description in the life sciences



Foreword

Table of Content

Introduction

1 Scope

2 Normative references

3 Terms and definitions

4 Criteria for formats and identifiers

5 Technical criteria and requirements

6 Semantic criteria and requirements

7 Requirements for ontologies suitable for annotation of biological data

8 Requirements for domain specific data standards

9 Requirements for data repositories for biological data

Annex A (informative) Recommended formats for life science data

Annex B (informative) Minimal reporting standards for data, models and metadata

ISO/TC 276 Biotechnology WG 5 (Data Processing and Integration) works on a draft for a new ISO guideline standard for data in the life sciences:

Reference framework („hub“) standard for (non-ISO) community standards

- Requirements and rules for the concerted application of community standards for formatting, description and documentation of datatypes in the life sciences
- Catalogue of criteria and requirements for interoperable life science data formats and semantic data description standards

Standardized and Harmonized Data Sharing: ISO 20691 (Draft)

Requirements for data formatting and description in the life sciences



FAIRsharing.org standards, databases, policies

Search all of FAIRsharing

Standards Databases Policies Collections Add/Claim Content Stats Log in or Register

bsg-c000079 Actions

Information: This record is currently having edits to it reviewed by FAIRsharing Admins...

This record is undergoing active curation and therefore the values may change.

TECHNICAL COMMITTEES
ISO/TC 276 ISO/CD 20691 Collection - DRAFT
Biotechnology

ISO is a worldwide federation of national standards bodies. The ISO Technical Committee ISO/TC 276 has a set of Working Groups (WG) working on standardization in the field of biotechnology processes; and WG5 focuses on Data Processing and Integration. The out put of this group is the ISO/CD 20691 specification that details the requirements for the consistent formatting and documentation of data and metadata in the life sciences and biotechnology, including biomedical research and non-human biological research and development; it covers manual or computational workflows. This Collection includes the standards detailed in the ISO/CD 20691 specification, and serves as a 'live' list to search and discover these standards, their use by repositories, as well as their evolution over time.

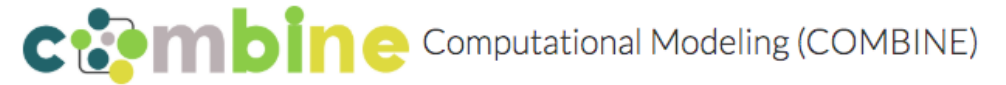
This record has no maintainer.

Record added: Jan. 28, 2021, 8:48 p.m.
Record updated: Feb. 1, 2021, 4:22 p.m. by [The FAIRsharing Team](#).

<https://fairsharing.org/collection/ISOCD20691CollectionDRAFT>



COMBINE Community Standards for Computational Modelling in Biology



doi:10.25504/fairsharing.9qv71f

Systems Biology Markup Language
Abbreviation: SBML

General Information

The Systems Biology Markup Language (SBML) is a machine-readable exchange format for computational models of biological processes. Its strength is in representing phenomena at the scale of biochemical reactions, but it is not limited to that. By supporting SBML as an input and output format, different software tools can operate on the same representation of a model, removing chances for errors in translation and assuring a common starting point for analyses and simulations.

Homepage <http://sbml.org>
Countries that developed this resource Worldwide
Created in 1999
Taxonomic range All

Knowledge Domains

Enzymatic Reaction Mathematical Model Molecular Entity Network Model Pathway Model

Subjects

Life Science Systems Biology

In the following recommendations:

EMBOpress Genetics & Genomics Next

How to cite this record FAIRsharing.org: SBML: Systems Biology Markup Language; DOI: <https://doi.org/10.25504/FAIRsharing.9qv71f>; Last edited: April 10, 2019, 10:49 a.m.; Last accessed: Dec 04 2019 4:52 p.m.

This record is maintained by [skeating](#) ORCID

Record added: May 14, 2015, 11:14 a.m.
Record updated: April 10, 2019, 10:47 a.m. by The FAIRsharing Team.

Show edit history

Legend

- DATABASE
- POLICY
- COLLECTION
- ▲ TERMINOLOGY ARTIFACT
- ◆ MODEL/FORMAT
- ◆ IDENTIFIER SCHEMA
- ◆ REPORTING GUIDELINE

Implementing Databases (17)

MetaCrop 2.0
The MetaCrop resource contains information on the major metabolic pathways mainly in crops of agricultural and economic importance. The database includes manually curated information on reactions and the kinetic data associated with these reactions. Ontology terms are used and publication identification available to ease mining the data.

Integrated Pathway Analysis and Visualization System
iPAVS provides a collection of highly-structured manually curated human pathway data, it also integrates biological pathway information from several public databases and provides several tools to manipulate, filter, browse, search, analyze, visualize and

Related Standards

Reporting Guidelines

Minimal Information Required In the Annotation of Models

Terminology Artifacts

Open Food Safety Model Ontology
Systems Biology Ontology

Models and Formats

CellML
Systems Biology Graphical Notation
Simulation Experiment Description Markup Language
Open Modeling EXchange format
Extensible Markup Language

<https://fairsharing.org/collection/ComputationalModelingCOMBINE>



Workshop Putting Science into Standard
Toward standardization", April 28-29, 20

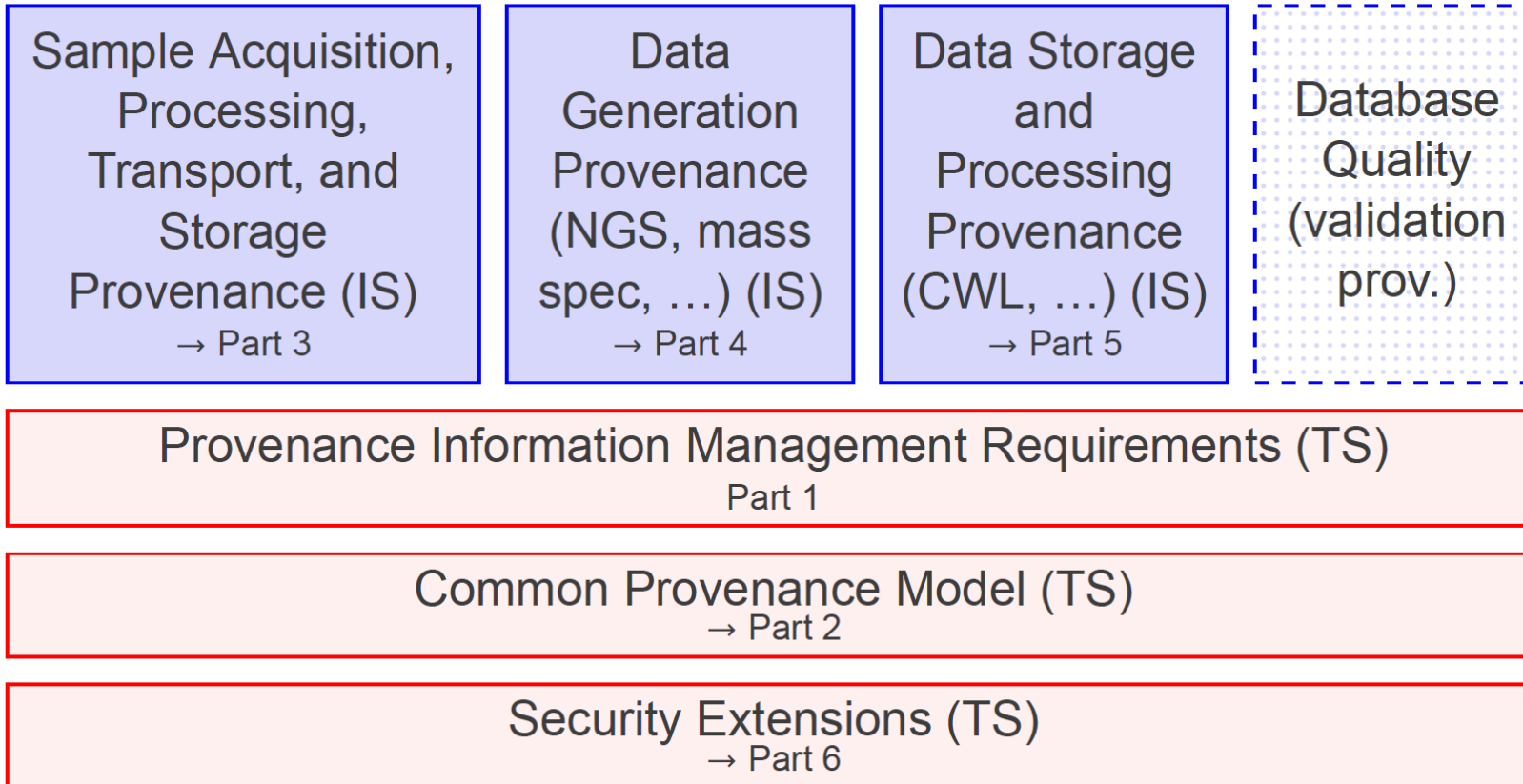


Standardized Data Provenance Information: ISO 23494 Series (Draft)

Provenance information model for biological specimen and data



EXPECTED STRUCTURE



- **history of biological samples** (including acquisition, processing, transportation, storage, and retrieval, etc.)
- **data history** (including generation of certain datatypes, processing, storage and validation)
- Based on W3C Prov

ISO/TC 276 Biotechnology



中国国家标准化管理委员会
Standardization Administration of the P.R.C.

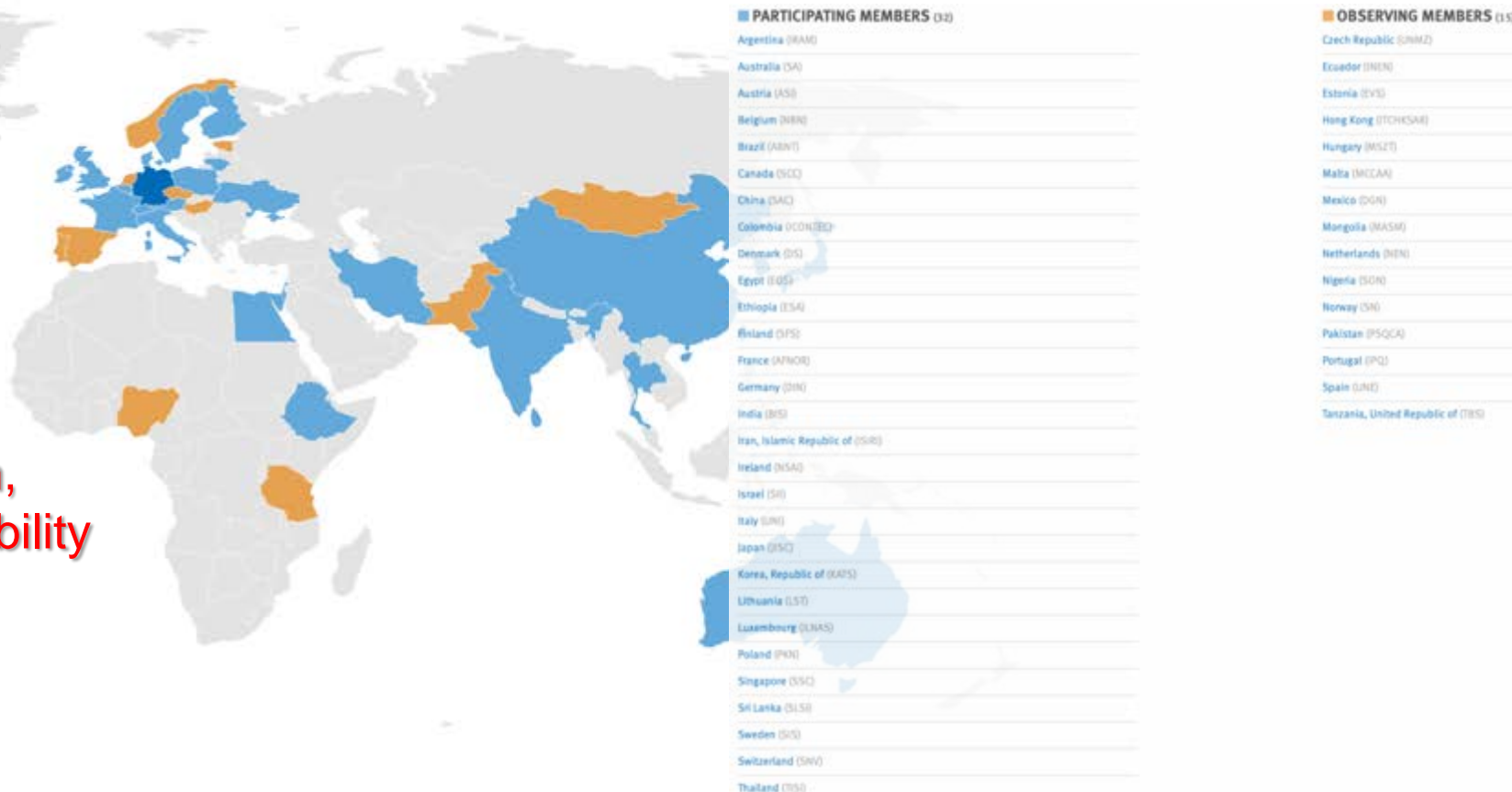


<http://www.iso.org/>

Scope:

Standardization in the field of biotechnology processes that includes the following topics:

- Terms and definitions
- Biobanks and bioresources
- Analytical methods
- Bioprocessing
- Data processing, annotation, analysis, validation, comparability and integration



EU-STANDS4PM: European standardization framework for data integration and data-driven *in silico* models for personalised medicine

H2020 projects
Use cases



Mission:

- › Establishing a pan-European Expert Forum to tackle the complexity of big data integration for *in silico* methodologies in personalised medicine

Aims:

- › Assessment of national strategies for data-driven *in silico* modelling approaches
- › Development of cross-border standards, recommendations and guidelines for *in silico* methodologies applied in personalised medicine

Key features:

- › Harmonisation of health/disease data integration strategies across Europe
- › Strengthening data-driven *in silico* approaches
- › Advise on health data integration and standards for research and industry
- › Open network that seeks interaction with all relevant stakeholders

Outcomes:

- **Harmonized Data Access Agreement (hDAA)** for Controlled Access Data
- **Legal and ethical review** of data integration in personalized medicine
- **Survey of data sources and models** in personalised medicine in Europe
- Brunak S, *et al.*, Journal of Integrative Bioinformatics, 17(2-3), 20200006

Standardization



Data and models



Legal/ethical frame



Regulator Coordinator



Workshop Putting Science into Standard “Organ-on-chip: Toward standardization”, April 28-29, 2021



Survey of data sources and models in personalised medicine in Europe

- Online survey with 92 questions
- 71 respondents (11 EU countries, UK and US)

Large variety in type of datasets/studies

- from 100 to 27×10^6 individuals included
- ICD10 & ICD9 most commonly used standards
 - others include ACT, SNOMED-CT, HL7, DICOM, Plink, SNPtest
- Sex and date of birth most common demographic data
 - Others included sometime: ethnicity, place of birth, socioeconomic status
- Genotypes and sequence data most common type of biological data followed by expression and epigenetics
 - Microbiome, metabolomics and proteomics rarely included

- Metadata for biological data not so often captured
 - Data source, type of tissue <70%
 - Methodology of data generation < 60%
 - Quality control procedure < 50%
 - Date of sampling <50%
 - Pre analytic steps < 40%
- Medication not that often captured, ATC standard not that commonly used
- Central data for personalised medicine such as response to treatment and adverse events are not capture so often



Ali Manouchehrinia
Karolinska Insitutet



Arshiya Merchant
Elixir



Niklas Blomberg
Elixir



Ingrid Kockum
Karolinska Insitutet

Topics to be discussed in this session

- **What to achieve with data standardization in OoC?**
 - Data integration across technology "silos" via interoperability/interfaces of (meta-)data standards for specific datatypes
 - Device interoperability via data exchange between systems
 - Establishing data analysis workflows based on harmonized interfaces ...
- Different approaches: Formal data standards (ISO, CEN,...) vs. community data standards driven by scientific initiatives
- **Technologies and datatypes in OoC that need formatting standardization?**
e.g. biophysical data, cell culture data, microenvironment data, physiological data, toxicology data, (bio)chemical data, biological sample data, bioactivities, medical data, OMICS data (genomics, proteomics, metabolomics,...), imaging data, models and simulation data, sensor data, ...
- Need for **data provenance standards** for traceability of the data and biological material
- Need for **metadata standards** (including terminologies) and for data **quality standards**
- **Interoperability** of subdomain-specific (meta-)data standards
- **Data access** for controlled data (e.g. person-related human data that falls under GDPR regulation)
- **Acceptance of data standards** by the OoC community
- **Standardization gaps** that need to be filled: e.g. Data integration standards, Data Provenance, Data Interoperability, AI technologies (data input/output)
- Which datatypes/fields have highest **priority**, which lower priorities to be standardized?

Thank you !

Martin Golebiewski, HITS
Heidelberg Institute for
Theoretical Studies, Germany



<https://www.eu-stands4pm.eu>

martin.golebiewski@h-its.org

