Standards for

Data acquisition and management

Regulatory & data reporting thematic area

Martin Golebiewski, HITS
Heidelberg Institute for Theoretical Studies, Germany

EU-STANDS4PM
standards for in silico models for personalised medicine

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

28th-29th April 2021

HITS

LiSyM
Liver Systems Medicine

nfdi 4 health

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

SUBJECT CATEGORIES
» Research data
» Publication characteristics

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

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

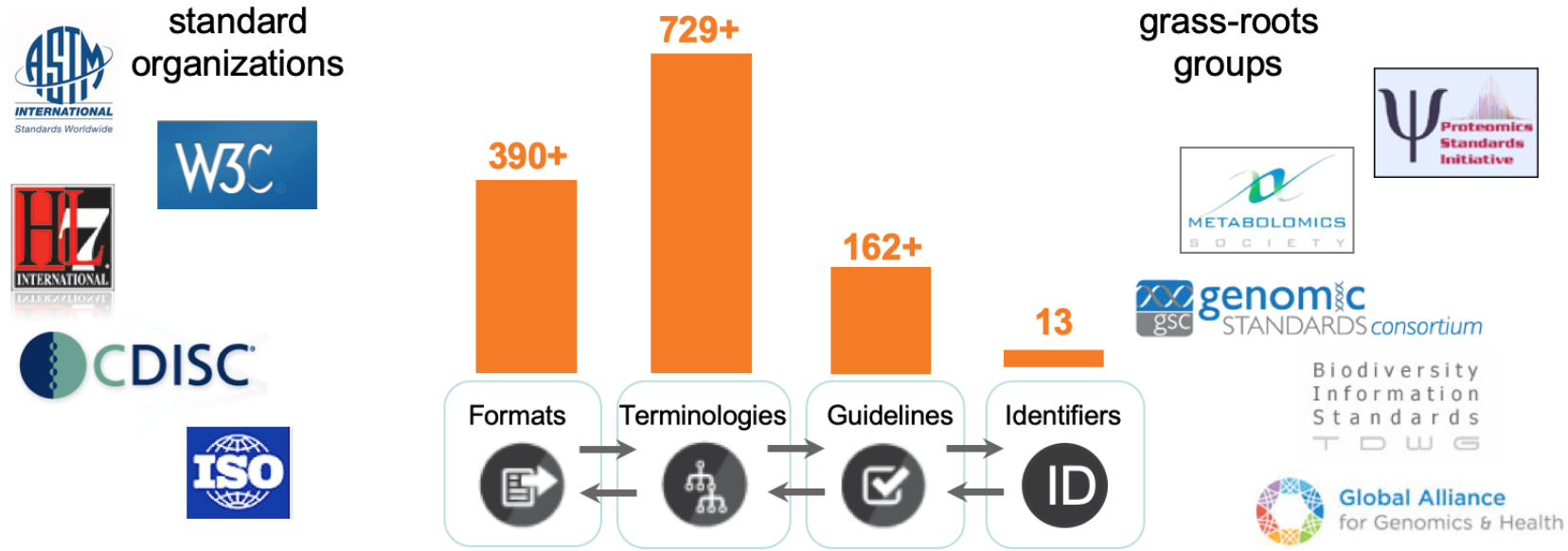
To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even 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

The Forrest of Standards in the Life Sciences



For modelling standards see review:

Golebiewski M: **Data Formats for Systems Biology and Quantitative Modeling**,
In: Ranganathan S., Nakai K., Schönbach C. and Gribskov M. (eds.) **Encyclopedia of Bioinformatics and Computational Biology**, Volume 2, 2019, Pages 884-893

Registry for Modeling Standards:

https://normsys.h-its.org

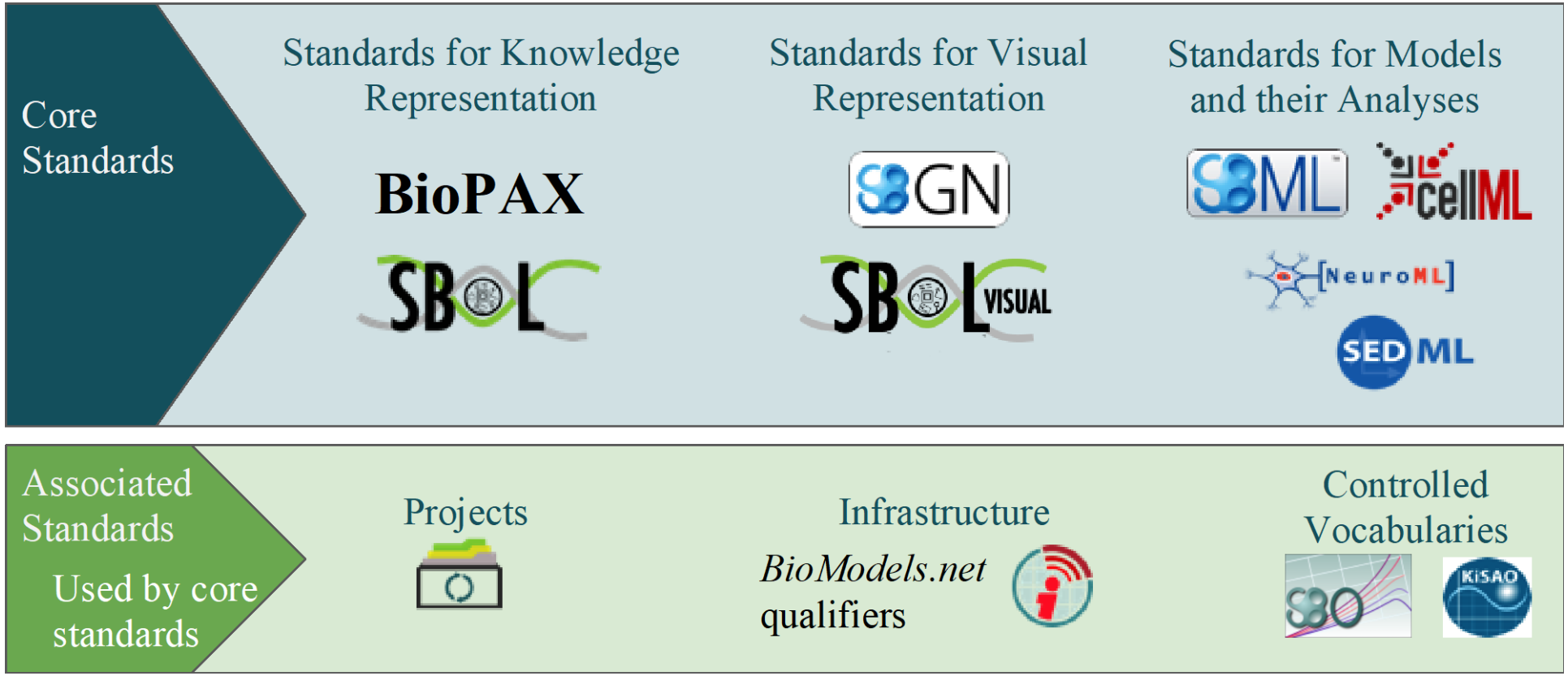
FAIRsharing.org
standards, databases, policies

Source: Susanna-Assunta Sansone (University of Oxford, UK)

COMBINE Community Standards for Computational Modelling in Biology

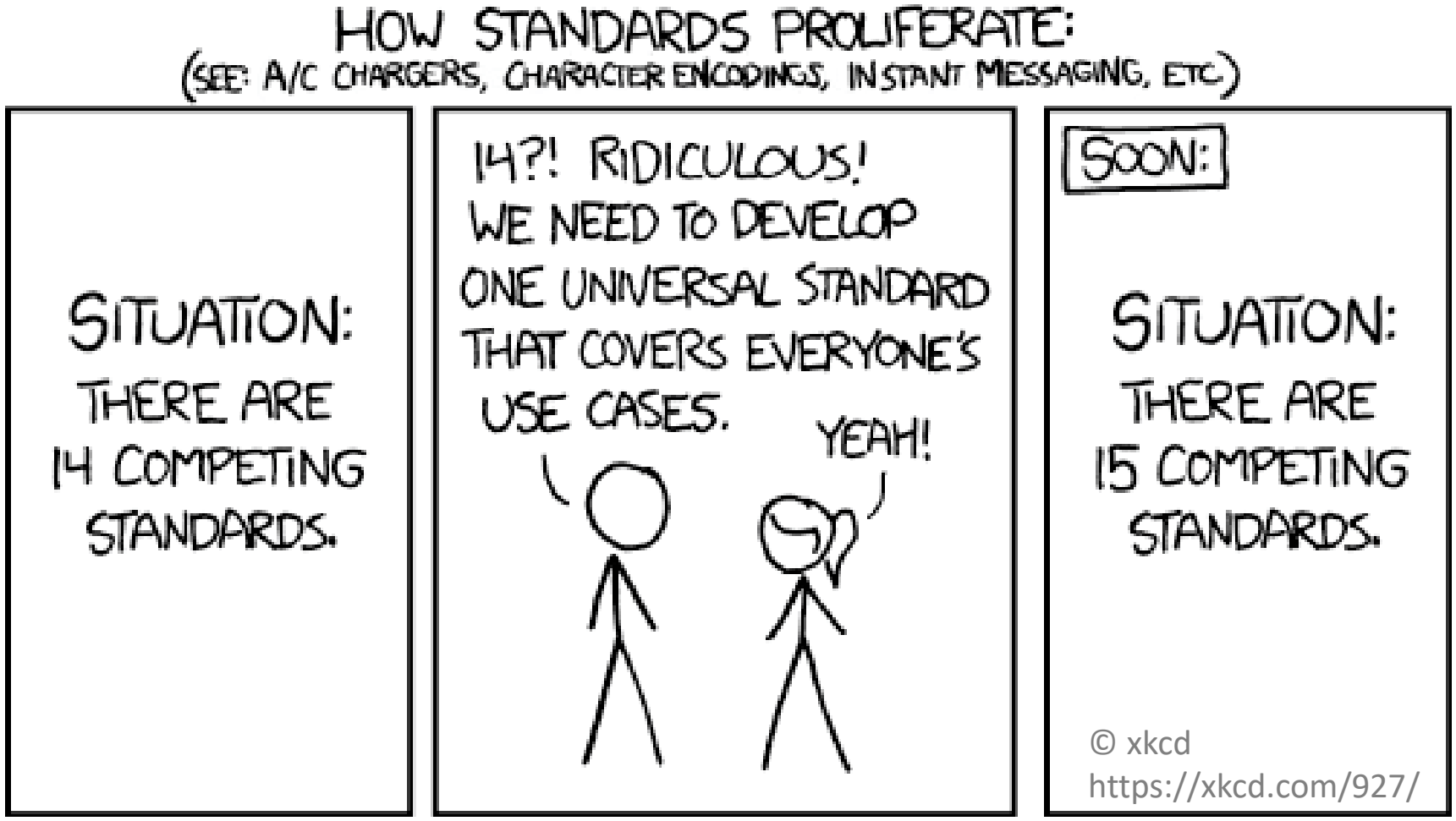
combine

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

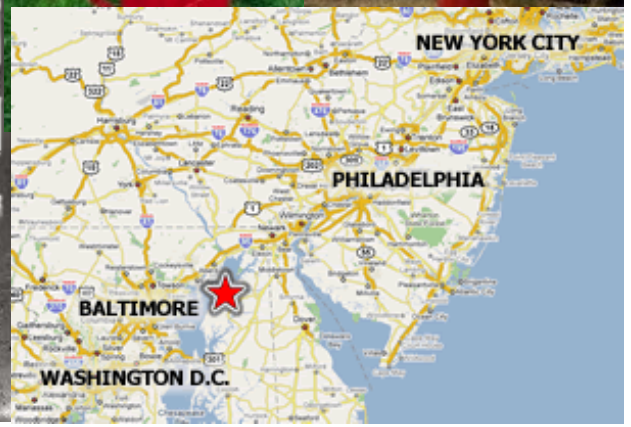
... so many standards



© xkcd
https://xkcd.com/927/

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

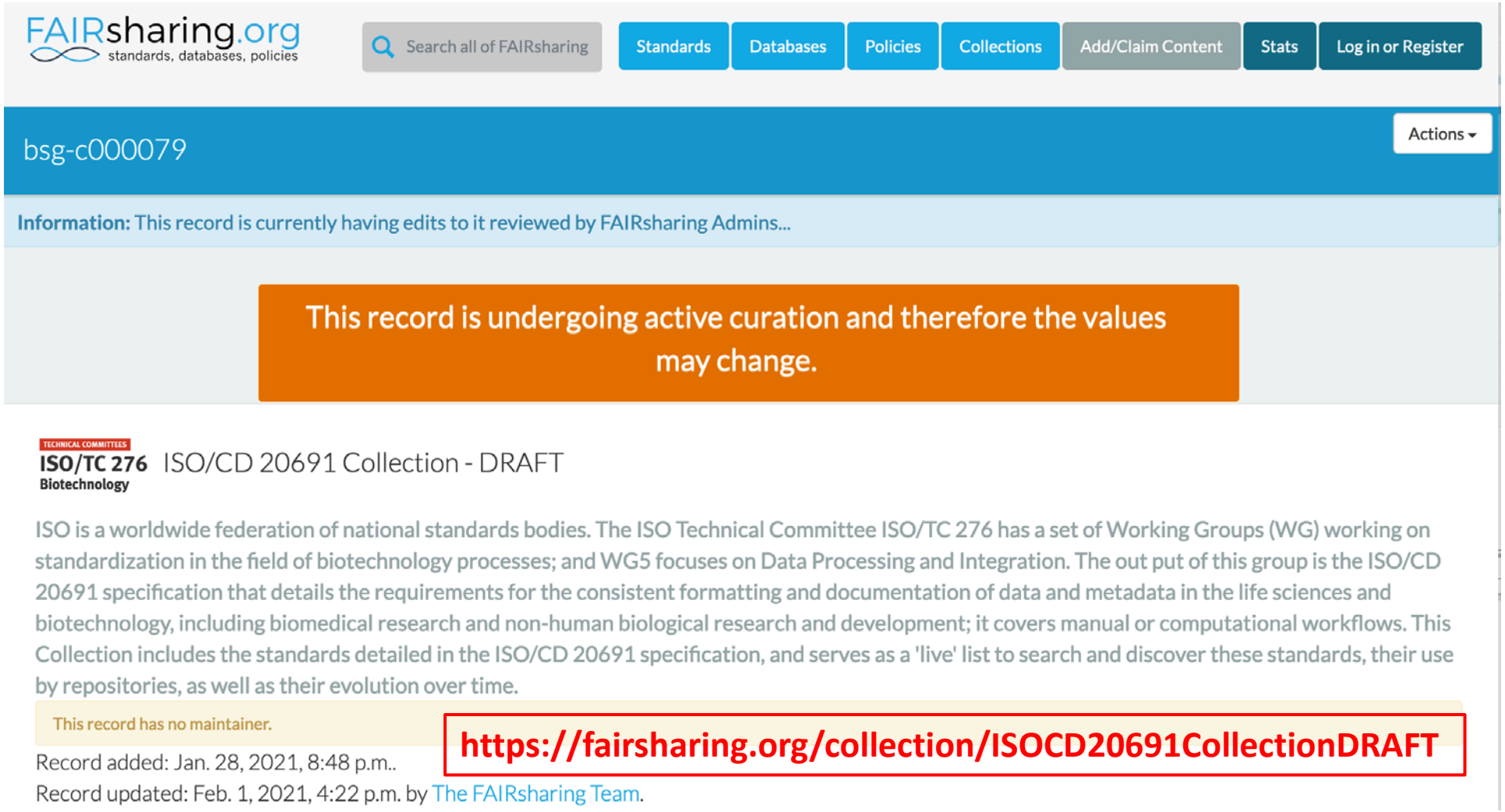
Table of Content

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



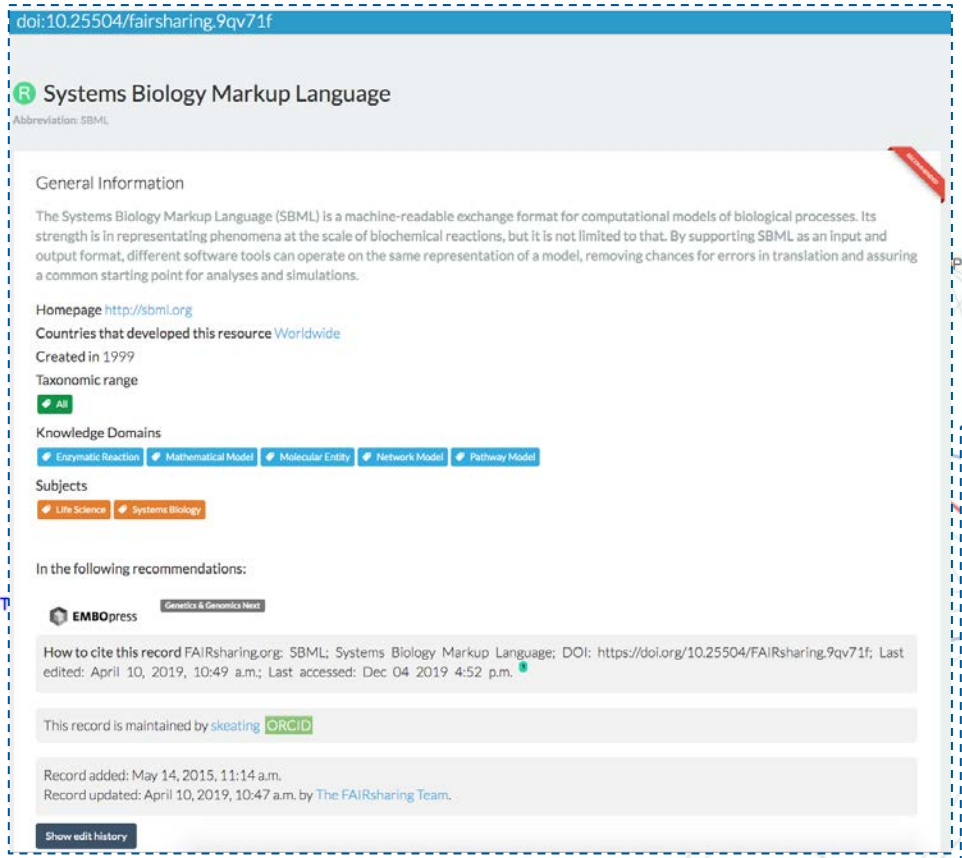
FAIRsharing.org Collections

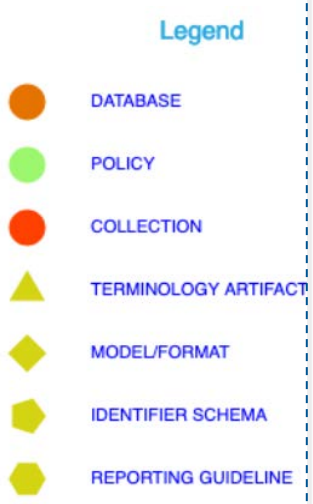
https://fairsharing.org/collection/ISOCD20691CollectionDRAFT

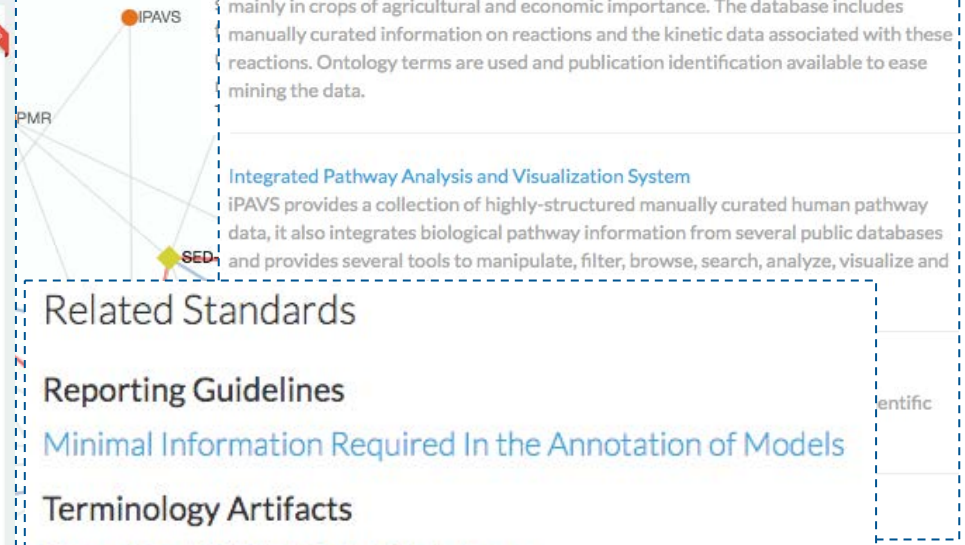
COMBINE Community Standards for Computational Modelling in Biology

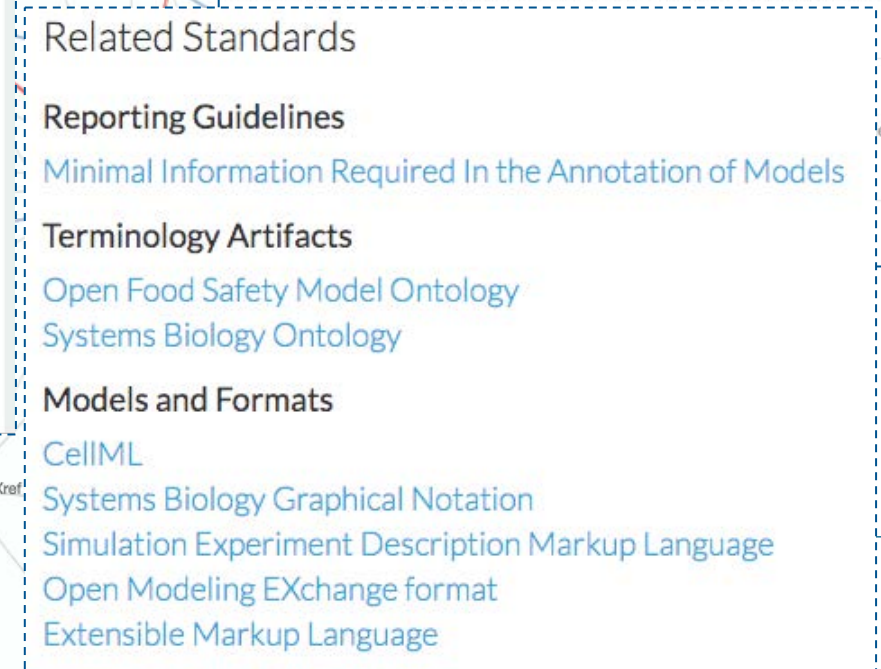
FAIRsharing.org Collections

combine Computational Modeling (COMBINE)









https://fairsharing.org/collection/ComputationalModelingCOMBINE

Standardized Data Provenance Information: ISO 23494 Series (Draft)
Provenance information model for biological specimen and data

EXPECTED STRUCTURE

Sample Acquisition, Processing, Transport, and Storage Provenance (IS) → Part 3	Data Generation Provenance (NGS, mass spec, …) (IS) → Part 4	Data Storage and Processing Provenance (CWL, …) (IS) → Part 5	Database Quality (validation prov.)

Provenance Information Management Requirements (TS)
Part 1

Common Provenance Model (TS)
→ Part 2

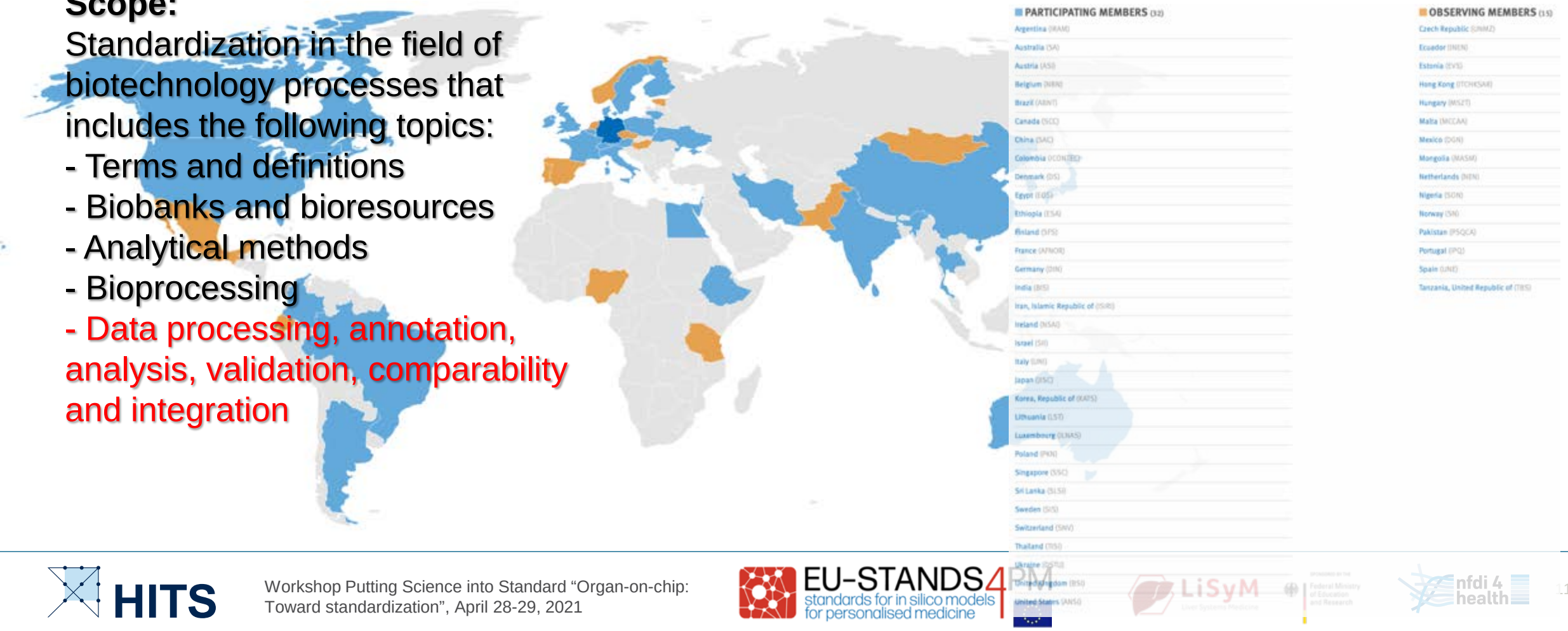
Security Extensions (TS)
→ Part 6

- **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

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

UCL

UNIVERSITÀ DI PARMA — BD2DECIDE

QIAGEN — SPIDIA4P

C|A|U Christian-Albrechts-Universität zu Kiel — SYSCID

Rotterdam University — LifeCycle

Karolinska Institutet — MultipleMS

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

Standardization

DIN — QIAGEN — HITS

Data and models

elixir — BAYER — University of Copenhagen — Universität Rostock

Legal/ethical frame

University of Copenhagen — University of Oxford — Universitas Vilnensis

Regulator

famhp

Coordinator

JÜLICH FORSCHUNGSZENTRUM

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

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/interfacing 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

EU-STANDS4PM
standards for in silico models for personalised medicine

https://www.eu-stands4pm.eu

martin.golebiewski@h-its.org

HITS

LiSyM
Liver Systems Medicine

nfdi 4 health