Standards for virtual human twins

and their implementability

Gerhard Mayer and Martin Golebiewski

Heidelberg Institute for Theoretical Studies

EDITH

EDITH project has received funding from the EU H2020 Research and Innovation Programme, under Grant Agreement n. 101083771

Agenda of the Standards Break-out session

Presentation:

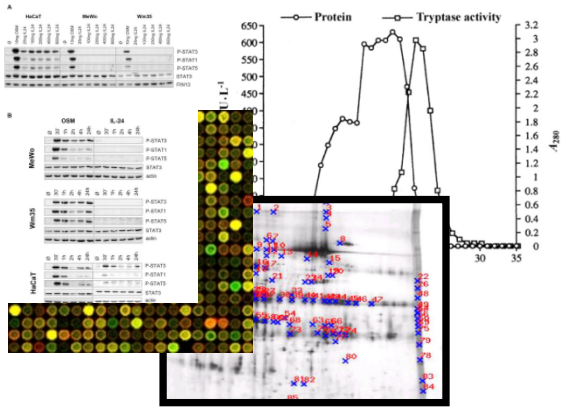
- Standardization for the VHT: Topic introduction (Martin)
- EDITH FAIRsharing collection demo (Martin)
- Introductions to the EDITH implementation guide (Gerhard)
- Introductions to the EDITH standards document (Gerhard)
- Discussion (Questions, Remarks, Comments, Ideas, …)

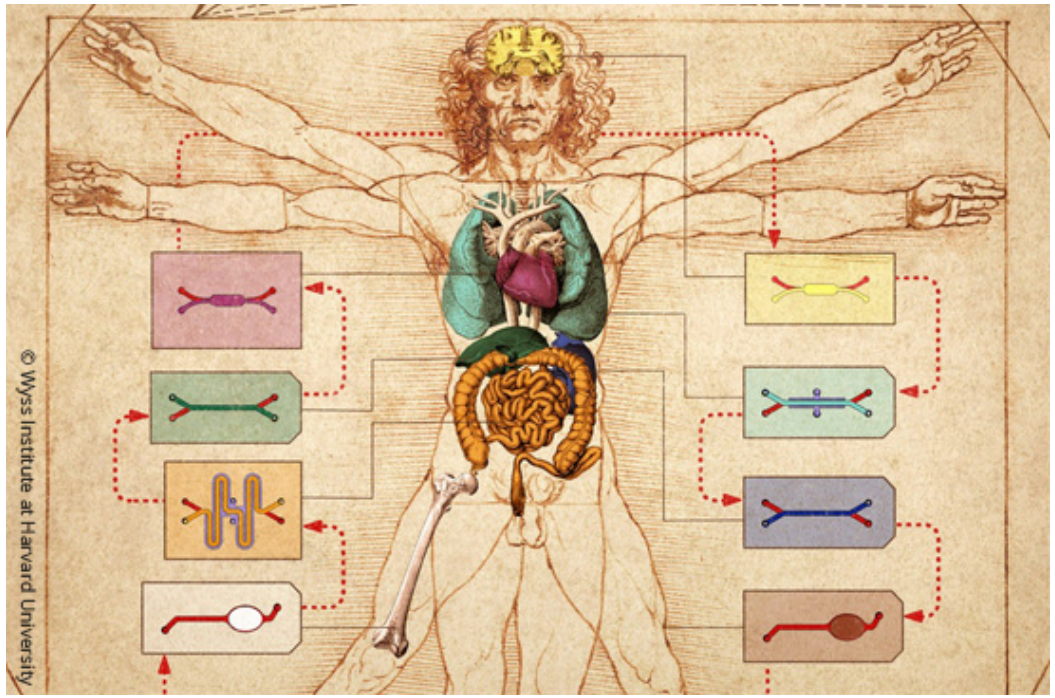
Questions to the audience:

- Are the presented standards feasible and suited for the implementation in the EDITH infrastructure?
- Are there missing standards, terminologies or standardization services?
- Which minimum metadata do we need to semantically describe a VHT?

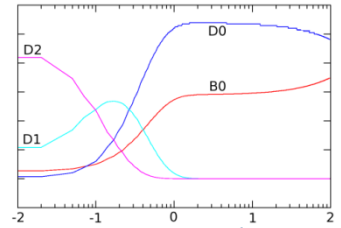
Data Integration for the VHT

Data from the donor



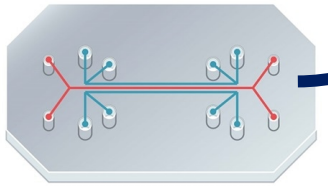


Simulation

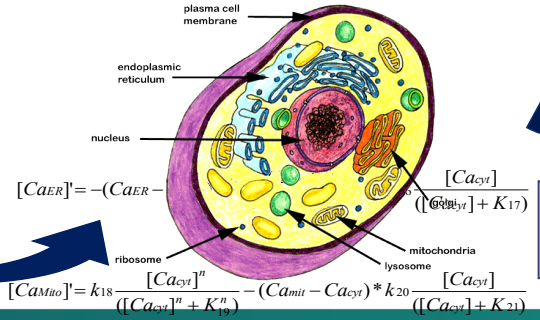


simulate

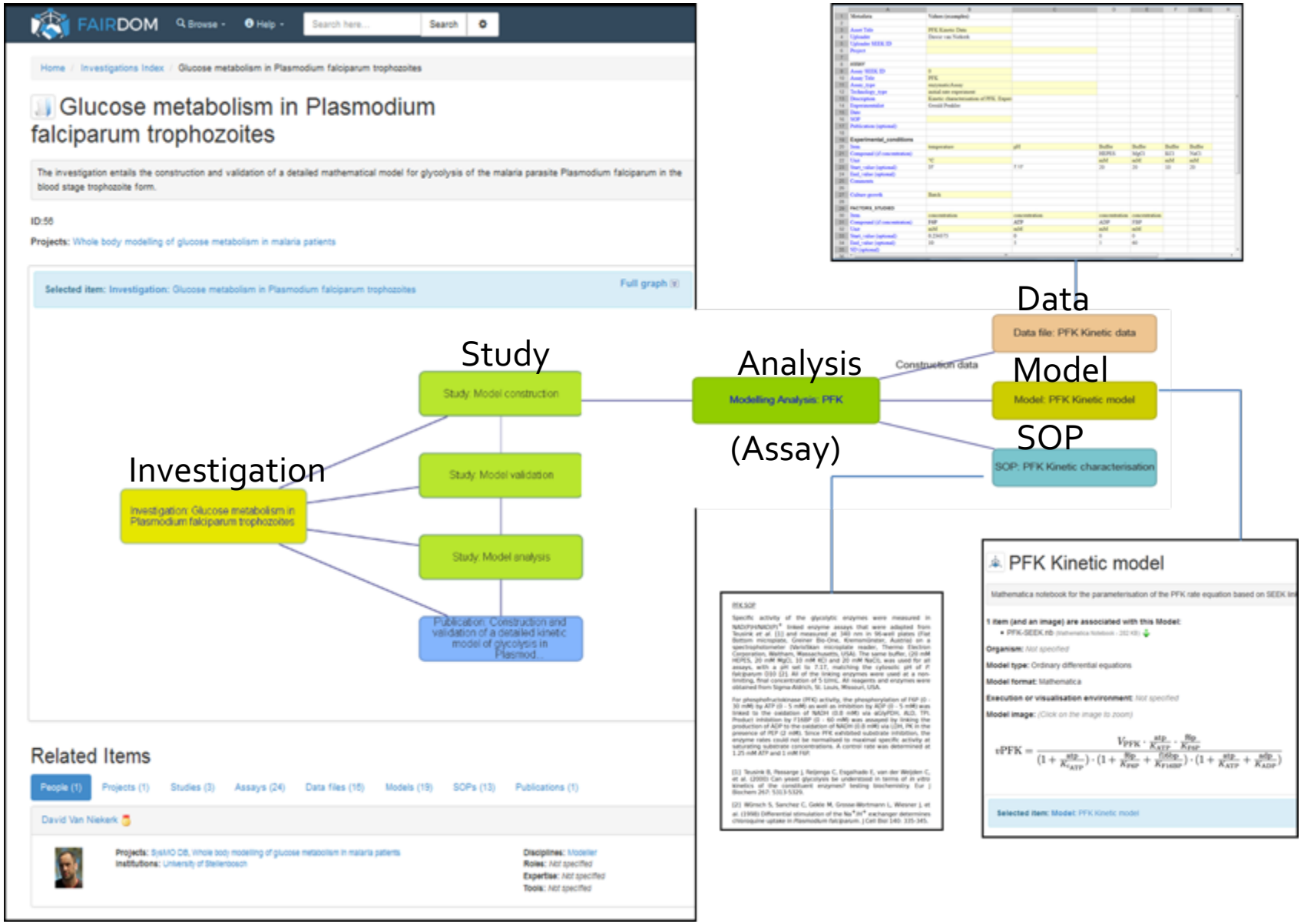
Process Data & SOPs



collect & integrate



Model

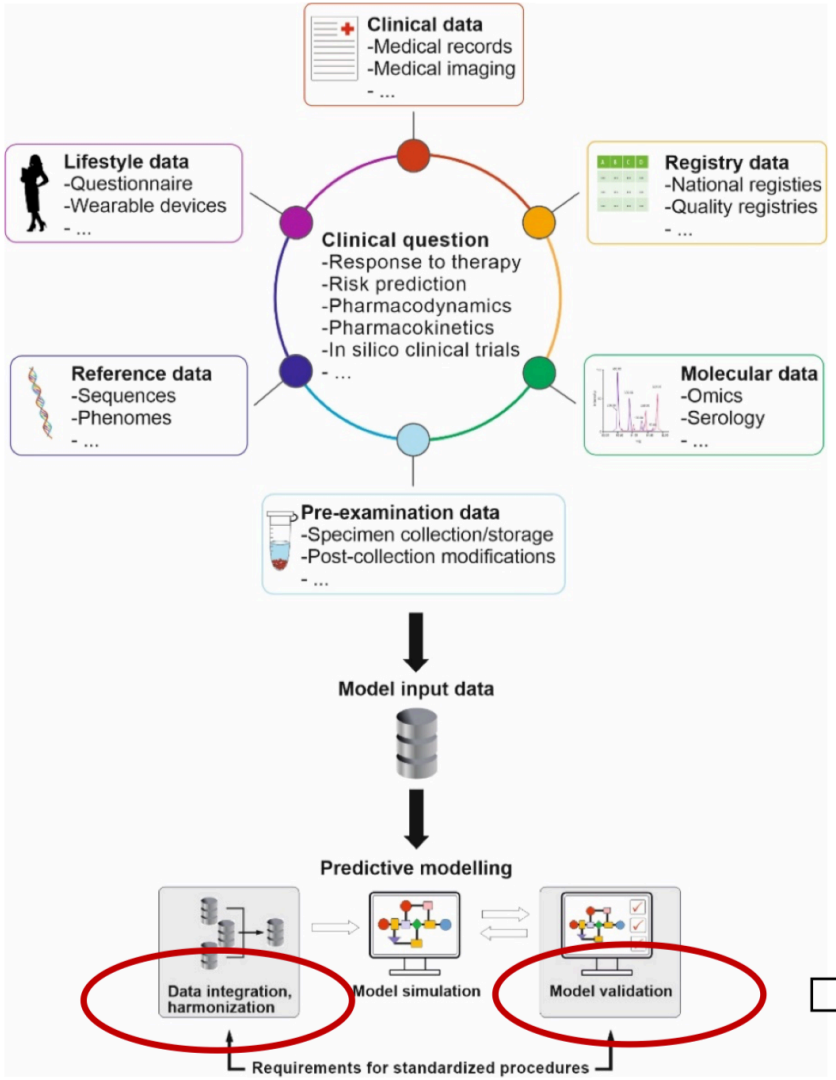


Integrative Data Management Systems: e.g. FAIRDOM SEEK

Wolstencroft K, Krebs O, Snoep JL, Stanford NJ, Bacall F, Golebiewski M, Kuzyakiv R, Nguyen Q, Owen S, Soiland-Reyes S, Straszewski J, van Niekerk DD, Williams AR, Malmström L, Rinn B, Müller W, Goble C: **FAIRDOMHub: a repository and collaboration environment for sharing systems biology research.** *Nucleic Acids Research 45(D1): D404-D407 (2017).* DOI: 10.1093/nar/gkw1032

Wolstencroft K, Owen S, Krebs O, Nguyen Q, Stanford NJ, Golebiewski M, Weidemann A, Bittkowski M, An L, Shockley D, Snoep JL, Mueller W, Goble C: **SEEK: a systems biology data and model management platform**. *BMC Systems Biology, 9: 33 (2015).* DOI: 10.1186/s12918-015-0174-y

Standardization Requirements for the VHT



Recommendations and standards for

- ⇨ Data integration
- ⇨ Model validation
- ⇨ Legal/ethical issues (e.g. patient rights, GDPR)

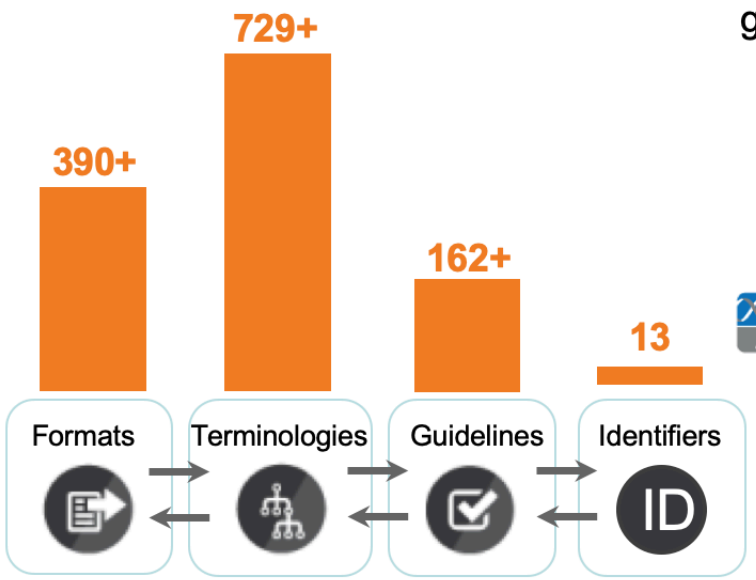
e.g.

- **ISO 20691:2022 Requirements for data formatting and description in the life sciences**
- **ISO TS 9491:2023 Biotechnology — Recommendations and requirements for predictive computational models in personalised medicine research** — Part 1: Guidelines for constructing, verifying and validating models
  Part 2: Guidelines for implementing computational models in clinical integrated decision support systems
- **ISO 23494 series: Biotechnology — Provenance information model for biological material and data**
- ISO 4454:2022 Genomics informatics — Phenopackets: A format for phenotypic data exchange
- ASME VV40: Assessing Credibility of Computational Modeling through Verification and Validation: Application to Medical Devices
- Community standards: e.g. COMBINE, GA4GH, etc.

The Forrest of Standards in the Life Sciences

standard organizations

grass-roots groups



COMMUNITY STANDARDS
for metadata and identifiers

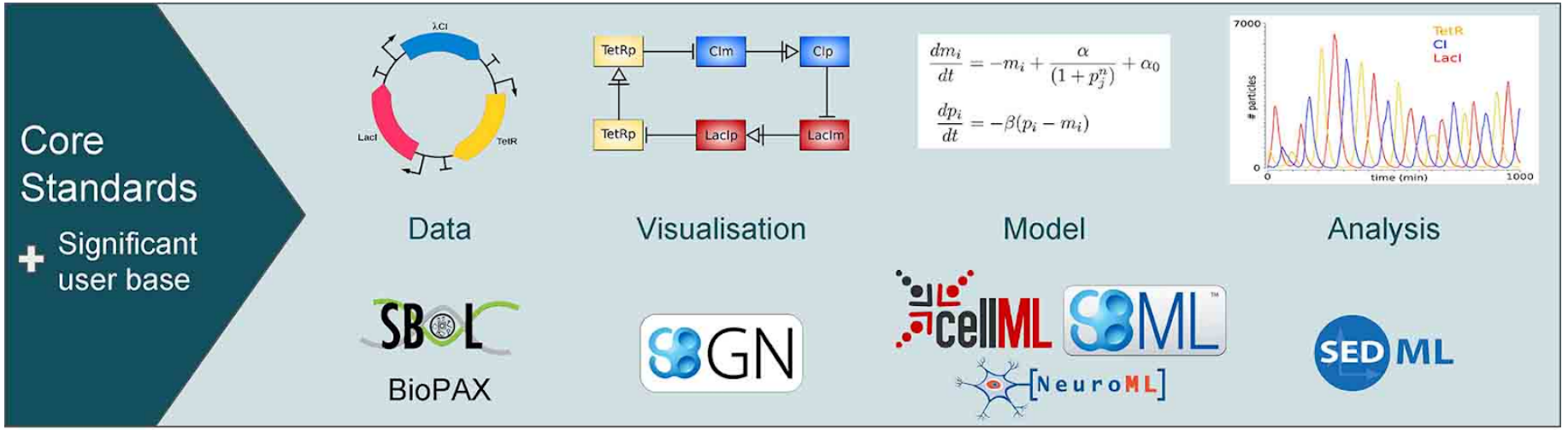
~1300

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

FAIRsharing.org
standards, databases, policies

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

COMBINE Community Standards for Computational Modelling in Biology



combine

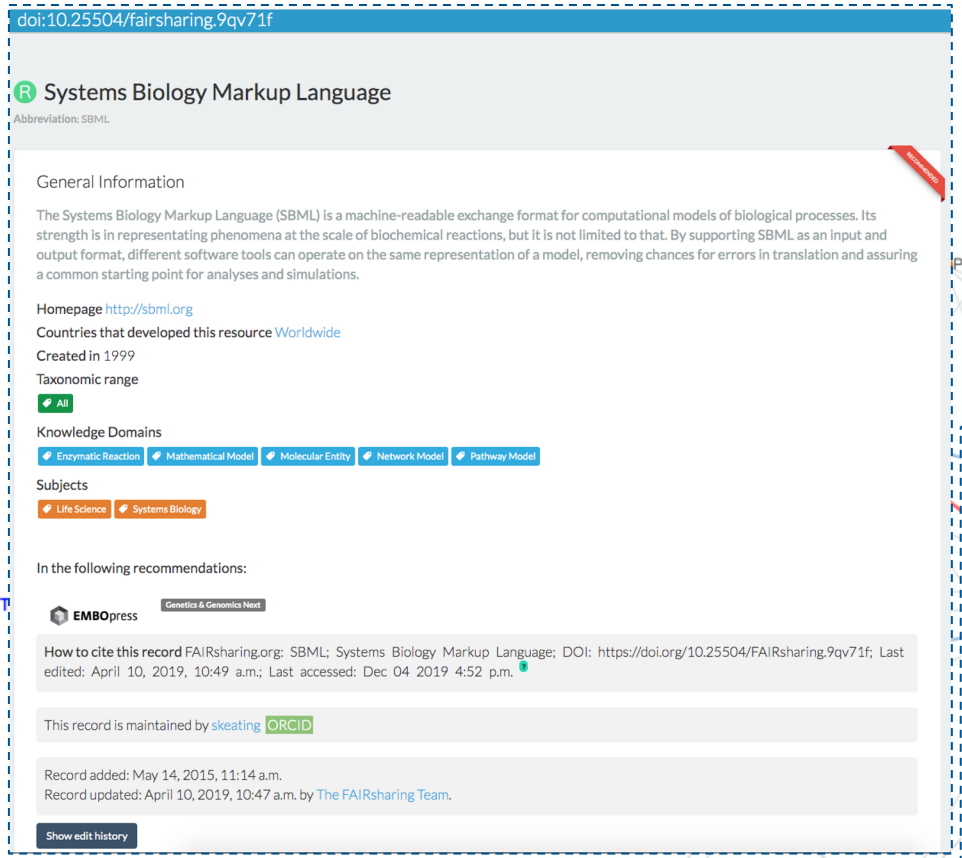
http://co.mbine.org

Schreiber et al (2020).
DOI:
https://doi.org/10.1515/jib-2020-0022

COMBINE Standards for Computational Modelling in Biology

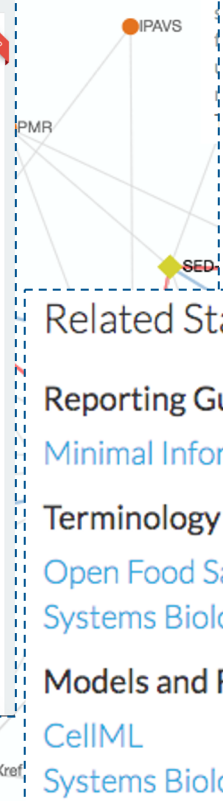
FAIRsharing.org Collections

combine Computational Modeling (COMBINE)



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

Standardized and Harmonized Data Sharing: ISO 20691:2022 Requirements for data formatting and description in the life sciences

Reference framework („hub“) standard for data 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

SCOPE:

- **Guidance** on rendering data in the life sciences findable, accessible, interoperable and reusable (FAIR)
- **Requirements for the consistent formatting and documentation of data and metadata** in the life sciences (experimental and procedural data obtained manually, as well as machine derived)
- **Requirements for storing, sharing, accessing, interoperability and reuse of data and metadata** in the life sciences
- Applicable to manual or computational workflows that systematically capture, record or integrate data and corresponding metadata
- Applicable to many domains in biotechnology and the life sciences (including the domains relevant for VHTs)

Standardized and Harmonized Data Sharing: ISO 20691:2022 Requirements for data formatting and description in the life sciences

6.2 Minimum consensus information for annotation of biological data

6.2.1 General

The minimum required annotation describes the contextual semantics of a data set, its parameters and results, including the biological, medical and environmental context. This metadata annotation should concisely describe both, the basic objective (e.g. problem addressed) of the process that produced the data set (e.g. the analytical or experimental procedure), and its context, components, independent (controlled, varied) quantities and dependent (emergent) measurables. These annotations can be reified as a simple table; converted into a standard semantic format, such as the Resource Description Framework (RDF)[9], or written to be compatible with the "Linked Open Data" concept of the W3C Data Activity[18]. The syntax for the annotations is a series of triplet phrases of the form subject – predicate – object. For example; "liver" – "is a(n)" – "organ". Table 2 lists examples of predicates that should be used for data annotations. The exact syntax and reification shall be "fit for purpose." For example, for data search and retrieval using standard web search engines the syntax shall be suitable for indexing by those search engines.

The required and suggested items for inclusion in the annotation are described below (see also Table 1 for examples of basic required biological descriptors). In addition, the suggested predicate for each item is given. For example, hepatocyte should be annotated with an URI pointing to the corresponding term or entry in a referenced resource, such as controlled vocabulary, domain ontology or terminology, using URIs of resolution services which guarantee the perennial resolvability of the URI where feasible (e.g. the URI referenced in [19], that points and refers to the corresponding term "hepatocyte" in the Foundational Model of Anatomy Ontology FMA[20]). Additionally, for being human readable it can be annotated with the respective common names "hepatocyte", "hepatic parenchymal cell", etc.

Table 1 — Examples of basic required biological descriptors

Field name (subject)	Predicate	Suggested ontology or vocabulary	Comments	Object – Human readable examples [a]
Species	is	NCBI Taxonomy[21]	The species the experiment was carried out in.	Human, *Escherichia coli*
Sex	is		The sex of the test subject, or of the source tissue or cells, where applicable.	male, female, male–female, female–male, hermaphrodite, and other applicable options
Age	is		Age of the individual the study was done in, or the age of the individual supplying the sample.	2 years, 8 h post fertilization (HPF)
Organ				
Tissue				
Cell				

Table 2 — Predicate (qualifier) examples

Predicate	Descriptions
is	The biological entity or process represented by the data set element has identity with the subject of the referenced resource. This predicate is used to link the component of the data set to its exact representation in another resource, controlled vocabulary or ontology; e.g. to link a hepatocyte cell in a data set to the term "hepatocyte" in an ontology.
isDescribedBy	The biological entity or process represented by the data set element is described by the subject of the referenced resource. This relation can be used, for instance, to link a species or a parameter to the literature that describes the concentration of that species or the value of that parameter.
hasPart	The biological entity or process represented by the data set element includes the subject of the referenced resource, either physically or logically. For example, this relation can be used to link a cell to the subcellular parts it encloses cell or to link the description of components of a multi-component protein complex.
isPartOf	The biological entity or process represented by the data set element is a physical or logical part of the subject of the referenced resource. This relation can be used to link a data set component to a description of the complex in which it is a part. For example, this relation can be used to link subcellular parts to the enclosing cell or to link the description of a component of a multi-component protein complex to the complex.
isVersionOf	The biological entity or process represented by the data set element is a version or an instance of the subject of the referenced resource. This relation can be used to represent, for example, the 'superclass' or 'parent' of a particular biological entity.
hasVersion	The subject of the referenced resource is a version, or an instance of the biological entity or

Standardized and Harmonized Data Sharing: ISO 20691:2022 Requirements for data formatting and description in the life sciences

Annex A
(informative)

Recommended formats for life science data

A.4 Formats for computer models of biological systems

A.4.1 CellML

CellML is a machine-readable, XML-based[23] model description and exchange format for computer-based mathematical models[85][86]. CellML is a description language to define models of cellular and subcellular processes. It defines lightweight XML constructs that group mathematical relationships within modules. The variables used in the mathematics are defined within each module, and connections between variables in different modules can be specified. CellML supports component-based modelling, allowing models to import other models, or subparts of models, therefore strongly encouraging their reuse and facilitating a modularized modelling approach. A CellML model typically consists of components, which can contain variables and mathematics that describe the behaviour of each component. The format provides means to reuse and group components into hierarchical structures. All entities (elements) carry an identifier and mathematical definitions are encoded using MathML. The mathematical model is considered to be the primary data, and biological context is provided by annotating the variables and equations with metadata using the Resource Description Format (RDF) [9] [87].

A.4.2 Systems Biology Markup Language (SBML)

SBML is a machine-readable, XML-based[8] model description and exchange format for computational models of biological processes[88][85]. Its strength is in representing phenomena at the scale of biochemical processes, but it is not limited to this only. The evolution of SBML proceeds in stages in which each "Level" is an attempt to achieve a consistent language at a certain level of complexity. Since SBML Level 3 the format is modular, with the core usable in its own right and packages being additional "layers" adding features to the core. By itself, SBML core is suited to representing such things as classical metabolic

Annex B
(informative)

Minimal reporting standards for data, models and metadata

B.2 Minimum information standards

Table 3 — List of minimum information standards

Acronym	Name	Description and Homepage
CIMR	Core Information for Metabolomics Reporting	Minimal requirements for metabolomics experiments http://www.metabolomics-msi.org
MIABE	Minimum Information about Bioactive Entities	Reporting requirements for the publication of data on one or a series of bioactive entities, such as pharmaceuticals and pesticides http://mibbi.sourceforge.net/projects/MIABE.shtml
MIACA	Minimum Information About a Cellular Assay	Information guideline and a modular Cellular Assay Object Model (CA-OM) that can cover the range of cellular assays possible and which is the basis for efficient data exchange http://miaca.sourceforge.net
MIAME	Minimum Information About a Microarray Experiment	Information needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment http://fged.org/projects/miame
MIAPAR	Minimum Information About a Protein Affinity Reagent	Guideline for experimentalists who wish to unambiguously describe protein affinity reagents and their protein targets

ISO 20691:2022 – Online ”live” Annex

Requirements for data formatting and description in the life sciences

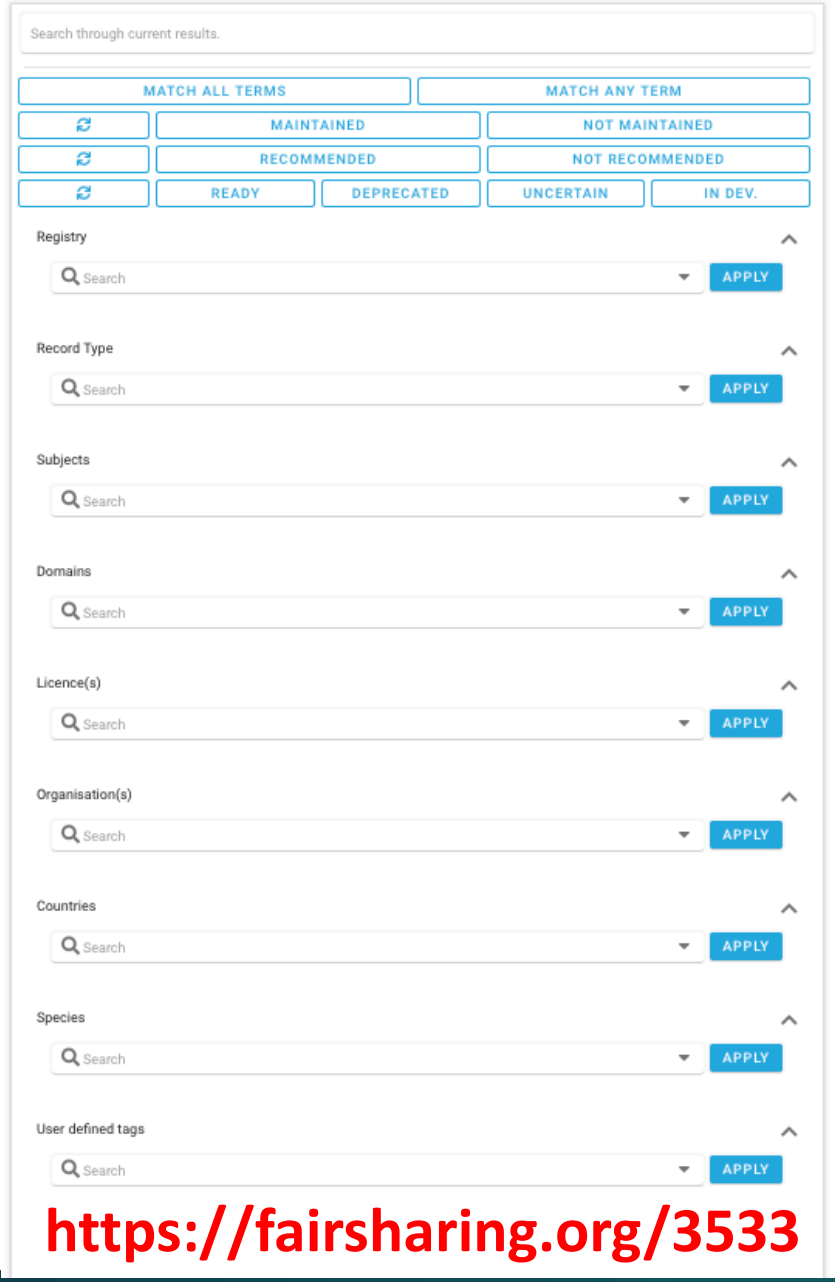
GENERAL INFORMATION

ISO 20691 - Requirements for data formatting and description in the life sciences

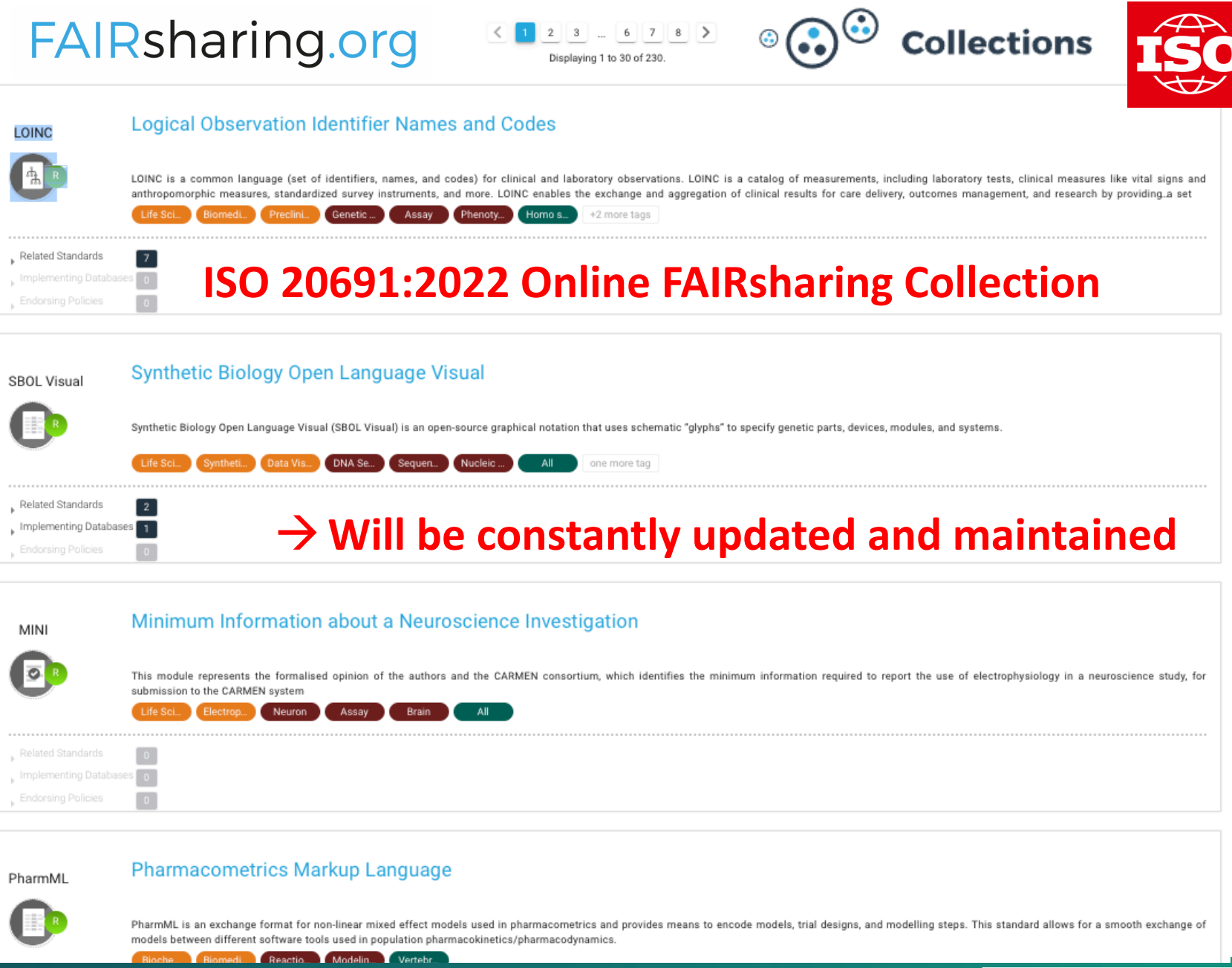
Type	Collection
Registry	Collection
Description	ISO is a worldwide federation of national standards bodies. The ISO Technical Committee ISO/TC 276 has a set of Working Groups (WG) working on standardization in the field of biotechnology processes; and WG5 focuses on Data Processing and Integration. One major ISO standard developed by this group is ISO 20691 Biotechnology — Requirements for data formatting and description in the life sciences. The document specifies requirements and rules 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 annexes of ISO 20691 as examples of common formats for life science data (annex A), as well as minimum reporting standards for data, models and metadata (annex B), and serves as a curated 'live' list to search and discover these standards, their use by repositories, as well as their evolution over time.
Organisations	Technical Committee 276, International Organization for Standardization (ISO), Geneva, Switzerland
Homepage	https://www.iso.org/committee/4514241.html
Reference URL	https://www.iso.org/standard/68848.html
Maintainers	Martin
Contacts	Martin Golebiewski
Subjects	Functional Genomics, Anatomy, Data Integration, Applied Microbiology, Epigenomics, Metagenomics, Biochemistry, Genomics, Medical Biotechnology, Bioinformatics, Clinical Studies, Data Management, Proteomics, Enzymology, Biotechnology, Agriculture, Computational Biology, Life Science, Glycomics, Metabolomics, Transcriptomics, Cell Biology, Phenomics, Database Management, Biomedical Science, Ontology And Terminology, Omics, Synthetic Biology, Systems Medicine, Epidemiology, Systems Biology, Medical Informatics
Domains	Mathematical Model, Reaction Data, Experimental Measurement, Biological Sample Annotation, Annotation, Sequence Annotation, Genome Annotation, Data Retrieval, Data Acquisition, Biobank, Cellular Assay, Modeling And Simulation, Omics Data Analysis, Biological Process, Small Molecule, Biological Sample, Mass Spectrometry Assay, Data Model, Sequence Alignment, Study Design, Phenotype, Pathway Model, Data Storage, Knowledge Representation, FAIR, Biocuration

FAIRsharing.org Collections

https://fairsharing.org/3533



https://fairsharing.org/3533



Ecosystem for Digital Twins in Healthcare

EU-STANDS4PM standards for in silico models for personalised medicine

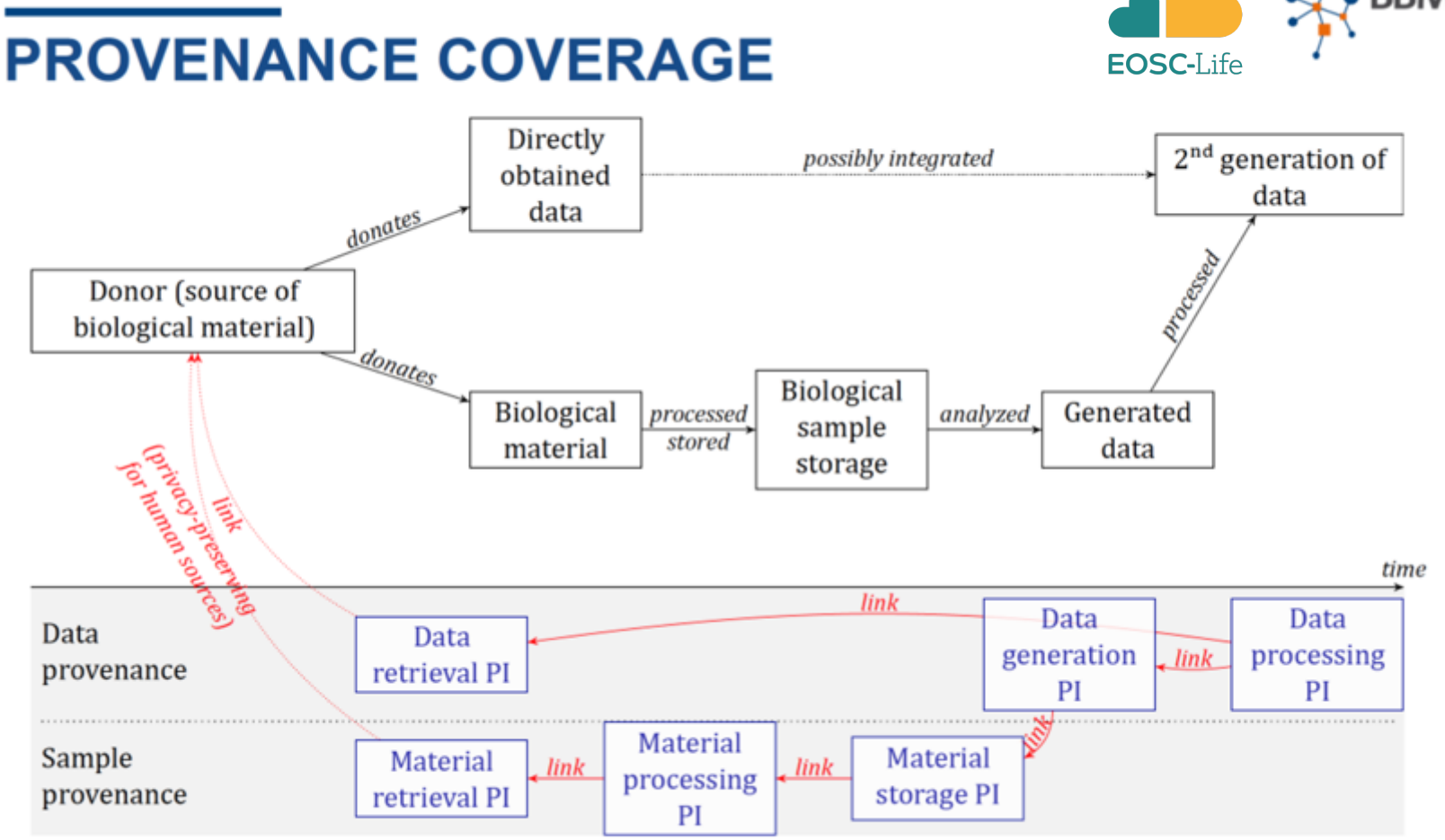
HITS

ISO 23494 series

Provenance information model for biological specimen and data

Scope:

- **History of biological samples:**
  - acquisition
  - processing
  - transportation
  - storage
  - retrieval ...
- **Data history:**
  - generation of datatypes
  - processing
  - storage
  - validation ...
- **Based on W3C Prov**



Standards for Model quality

- ASME VV 40: Assessing Credibility of Computational Modeling through Verification and Validation: Application to Medical Devices

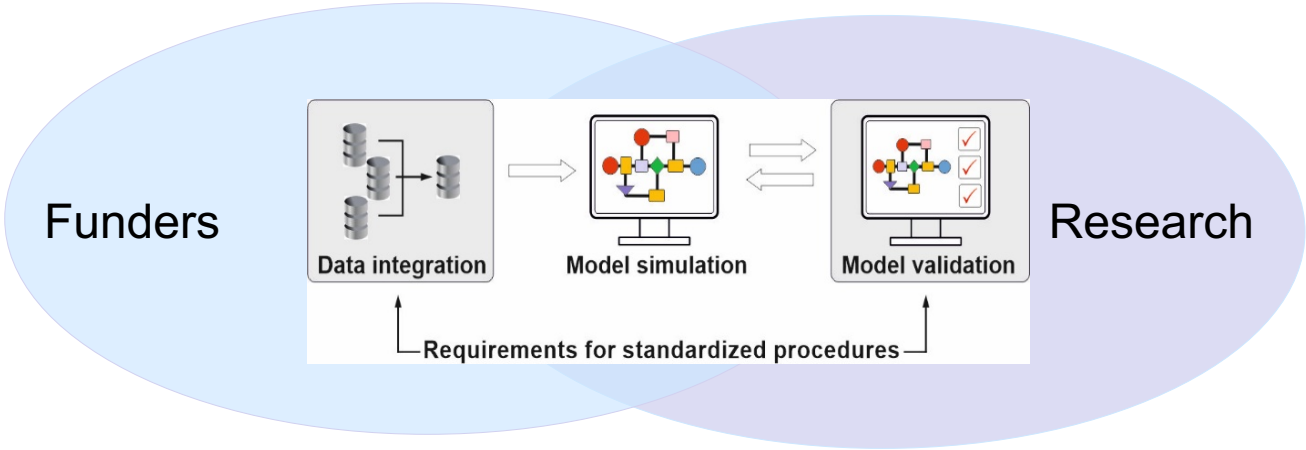
- ASME VVUQ SC 40 on Verification, Validation, and Uncertainty Quantification in Computational Modeling of Medical Devices

- GSP (good simulation practice initiative by Avicenna Alliance, VPHi & ISW project)

- ISO TS 9491-1:2023 Biotechnology — Recommendations and requirements for predictive computational models in personalised medicine research — Part 1: Guidelines for constructing, verifying and validating models (initiated by EU-STANDS4PM)

- ISO TS 9491-2 Biotechnology — Recommendations and requirements for predictive computational models in personalised medicine research — Part 2: Guidelines for implementing computational models in clinical integrated decision support systems (initiated by EU-STANDS4PM)

Rationale to develop an ISO Technical Specification for modelling in personalised medicine



Before ISO TS 9491:

- ⇨ No broadly accepted standards for health & disease data integration for predictive modelling
- ⇨ Many grass root but non-binding standards

ISO TS 9491-1 defines rules for

- ⇨ for data integration into models and VHT components
- ⇨ model setup, validation, simulation desciption & handling
- ⇨ adaptation of community standards for VHT and modelling processes in health and clinical research

ISO TS 9491 Biotechnology — Recommendations and requirements for predictive computational models in personalised medicine research — Part 1: Guidelines for constructing, verifying and validating models

Scope

This document specifies requirements and recommendations for:

- Design, development and establishment of predictive computational models
- Set-up, formatting, validation, simulation, storing and sharing of computational models
- Data used to construct or required for validating such models
- Formatting, descriptions, annotations, interoperability, integration, access and provenance of such data

for research purposes in the field personalised medicine.

Computational models used in routine clinical, diagnostic or therapeutic purposes are excluded.

ISO TS 9491 – Table of Content

Contents

ISO TS 9491 Content example: Models in personalised medicine

4.2.2 Cellular systems biology models

4.2.2.1 General

Simulation of complex dynamic biological processes and networks. Models can be either data-driven ("bottom-up") or mechanism-based ("top-down").

Mechanism-based concepts aim for a structural representation of the governing physiological processes based on model equations with limited amount of data, which are required for the base model establishment [16] or, alternatively, on static interacting networks [17, 18]. Data-driven approaches ([4], [19]) require, as the name implies, sufficiently rich and quantitative time-course data to train and to validate the model. Due to its often black-box nature, the model validation process in data-driven approaches relies on performance tests against known results.

4.2.2.2 Challenges

— Creation of models that balance the level of abstraction with comprehensiveness to make modelling efforts reproducible and reusable (abstraction vs size).

— Development of prediction models that can be adopted easily to individual patient profiles.

— Efficient parameter estimation tools to cope with population and disease heterogeneity.

— Overfitting of the model to the experimental/patient data and optimization methods for model predictions in a realistic parametric uncertainty.

— Flexibility in models to cope with missing data (e.g., diverse patient profiles).

— Scaling from cellular to organ and to organism levels (e.g., high clinical relevance, high hurdles for regulatory acceptancy).

ISO TS 9491 Content example: Data preparation

4.4 Data preparation for integration into computer models

4.4.1 General

Computational models in the life sciences in general and in personalised medicine research specifically are increasingly incorporating rich and varied datasets to capture multiple aspects of the modelled phenomenon. Data types are encoded in technology and subdomain specific formats and the variety and incompatibility, as well as lack of interoperability of such data formats have been noted as one of the major hurdles for data preparation.

To allow for seamless integration of data used for the construction of predictive computational models in personalised medicine these data shall:

- include or be annotated with sampling and specimen data that follow the recommendations and requirements as specified by the relevant domain-specific standards;
- be formatted using generally accepted and interoperable standard data formats commonly used for the corresponding data types (as specified by ISO 20691);
- include or be annotated with descriptive metadata that consider generally accepted domain-specific Minimum Information guidelines and describes the metadata attributes and entities using semantic standards, such as standard terminologies, controlled vocabularies and ontologies (as specified in Annex B of ISO 20691:20XX);
- follow best practice recommendations and requirements of generally accepted domain-specific data interoperability frameworks;
- structured in a way that allows integration of the data into a model, together with other data;
- include or be annotated with data provenance information that allows for tracking of the data and source material throughout the whole data processing and modelling;
- be made accessible via harmonized Data Access Agreements (hDAAs) for controlled access data, if open access to the data is not possible.

ISO TS 9491 Content example: Model formatting

4.5 Model formatting

4.5.1 General

If feasible (e.g. for mechanistic models), the encoding of a predictive computational model in personalised medicine shall be formatted using generally accepted, appropriate and interoperable standard model formats commonly used for the corresponding data types (as specified by ISO 20691 with extensive examples of standard formats recommended in annex A.3). The used data format and its version shall be unambiguously documented with the data to allow for later decoding of the data. The corresponding scientific communities have defined many grassroots standards to consistently structure and format data, models and their metadata for modelling [38] and simulation [31] in the life sciences. These standardization efforts are driven by standardization initiatives, such as the Computational Modelling in Biology Network (COMBINE)[17] ([39], [40]). For providing the potential users with an overview and comparable information about such standards, web-based information resources have been developed and are publicly available, such as the NormSys registry[18] for modelling standards.

Examples of established model formats mainly used for models consisting of molecular entities and describing their interactions and dynamical interplay include:

— SBML (Systems Biology Markup Language) as standardized interchange format for computer models of biological processes [41];

— CellML, a standard format to store and exchange reusable, modular computer-based mathematical models;

— NeuroML (Neural Open Markup Language), which allows standardization of model descriptions in computational neuroscience;

— SBOL for synthetic biology models;

— BioPAX for biological/biochemical pathway models [42].

ISO TS 9491 Content example: Sampling data

4.4.2 Sampling data

Dedicated measures need to be taken for collecting, stabilizing, transporting, storing and processing of biological specimen/samples, to ensure that profiles of analytes of interest (e.g. gene sequence, transcript, protein, metabolite) for examination are not changed ex vivo. Without these measures, analyte profiles can change drastically during and after specimen collection, thus making the outcome from diagnostics or research unreliable or even impossible, because the subsequent examination will not determine the situation in the patient, but an artificial profile generated during the pre-examination process. Important measures include for example times and temperatures of sample transportation not exceeding the specifications provided in ISO standards (ISO 20916:2019 as well as ISO 20186-1:2019) and ISO technical specifications (e.g. ISO/TS 20658:2017), giving guidelines on all steps of the pre-examination workflow. Conditions applied to specimen shall be documented in addition to other important metadata, including but not limited to the following Table 1.

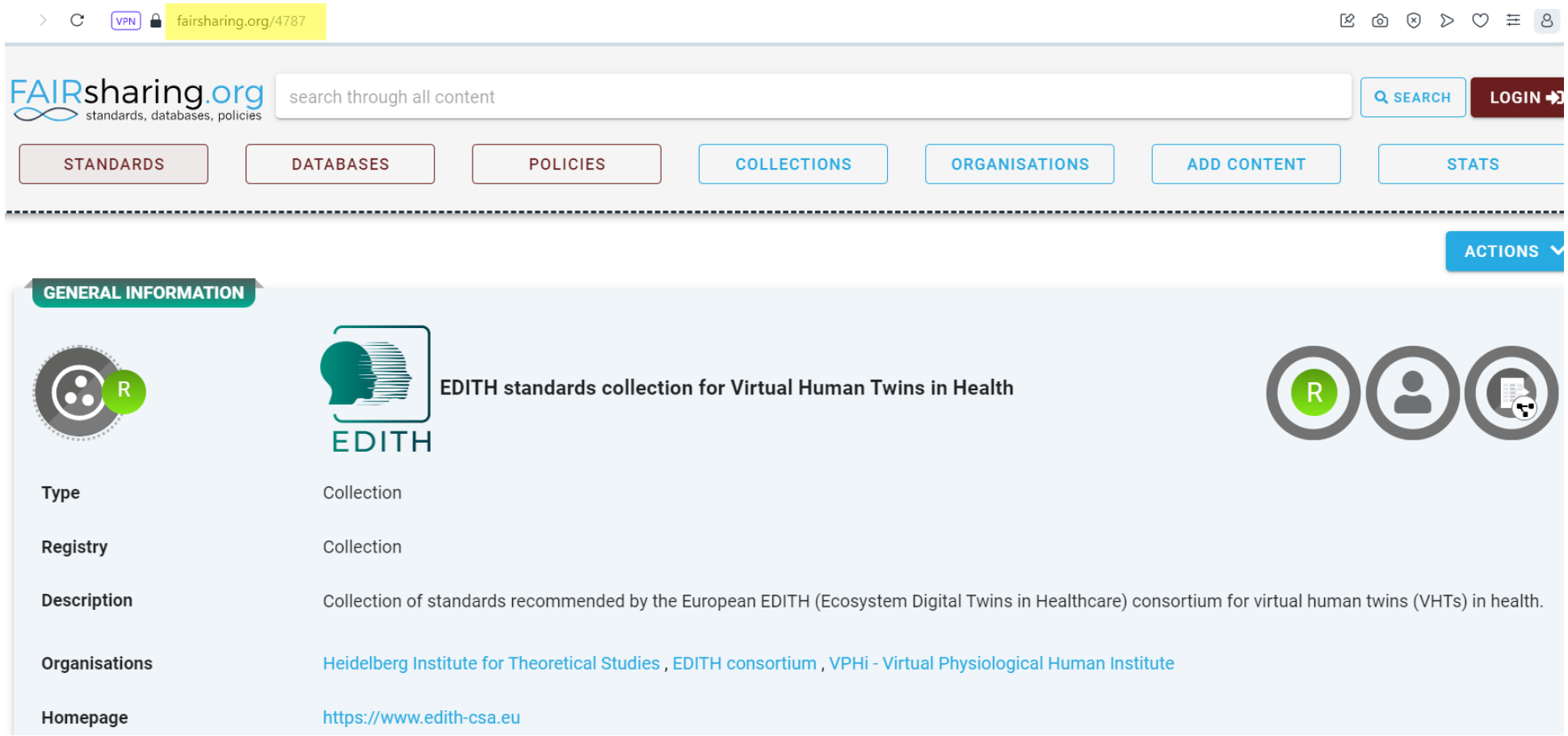
Table 1 — Examples of important metadata collected during pre-examination workflows

Metadata relevant in pre-examination processes	
Specimen collection	– ID of responsible person
Information about specimen donor	– ID – Health status (e.g. healthy, disease type, concomitant disease, demographics [e.g. age and gender]) – Routine medical treatment and special treatment prior to specimen collection (e.g. anaesthetics, medications, surgical or diagnostic procedures, fasting status) – Appropriate consent compliant with EU and national law from the specimen donor/patient
Information about the specimen, collection from the donor or patient and processing	– Type and the purpose of the examination requested – Specimen collection technique used (e.g. surgery, draw, flush) – Time and date when the specimen is removed from the body – Documentation of any additions or modifications to the specimen after removal from the body (e.g. addition of reagents)
Specimen storage and transport	– Temperatures of the collection device's surroundings
Specimen reception	– ID or name of the person receiving the specimen – Arrival date, time and conditions (e.g. labelling, transport

EDITH FAIRsharing collection

FairSharing collection (https://fairsharing.org/4787)

currently 153 standards, taxonomies and guidelines



Links to documents

- EDITH Standardization landscape, needs and gaps for the VHT
  https://zenodo.org/doi/10.5281/zenodo.10492795

- EDITH standards implementation guide
  https://doi.org/10.5281/zenodo.10524795

- FairSharing collection
  https://fairsharing.org/4787

Implementation guide (IG) and standards document

EDITH

EDITH project has received funding from the EU H2020 Research and Innovation Programme, under Grant Agreement n. 101083771

Implementation guide - Introduction

Resources:

- EDITH-T2.3 document "Analysing current landscape of standards, identifying needs and gaps"
- EDITH-T4.2 document "EDITH standards implementation guide (IG))"
- EDITH Fairsharing collection of standards, terminologies and guidelines (https://fairsharing.org/4787)

Best practices:

- Toward Good Simulation Practice (**GSP**) guidelines of the Avicenna Alliance: risk-informed credibility assessment of *in silico* simulation models
- Modelling Good Research Practices defined by **ISPOR** and **SMDM**

Recommended standards for Europe:

- European Electronic Health Record Exchange Format (**EEHRxF**)
- IEEE P7003 Standard for **Algorithmic Bias** Considerations
- IEEE P7001/D4 Draft Standard for **Transparency** of Autonomous Systems
- IEEE 7000 Standard Model Process for Addressing **Ethical Concerns** during System Design
- …

Implementation guide – Data handling

Data preparation (ISO 20691:2022 and ISO/TS 9491-1:2023)

- **sampling** the data.
- **data formatting and harmonization**, e.g., lab value concentrations must have a **unit** associated with them. This unit can either be mass/volume or mol/volume. Therefore, the values shall be converted to a unique scale. For that the molecular weight of the analyte must be known.
- data description by **descriptive metadata**, describing for example the context of the datasets.
- **semantic annotation** of the data, e.g., by annotating genes and proteins with ontology terms.
- definition of a data **interoperability framework** (e.g. BRIDG, FHIR, CDISC-OMOP, DCM, …)
- **data integration**, either on the personal or on the variable level → use of persistent identifiers.
- adding data **provenance information**.
- defining who can **access** the data.
- **→ FAIReR data**

Implementation guide - Schemas

Input and output interfaces should avoid **unstructured** .csv or .txt data

- Use of **schemas** for describing the data formats, e.g.
  - JSON schema
  - XML schema (.xsd)
  - ObjTables for describing the content of .csv files
  - Own clearly documented "schema" for .csv files uniquely describing the used cells (values and units)

- Standardized **serialization formats** for binary data
  - Apache Avro
  - Google protobuf

Implementation guide – Data integration and metadata

- Use of persistent identifiers like compact Uniform Resource Identifiers (**CURIEs**). General form "**prefix:local unique identifier**", where the prefix encodes the resource; Bioregistry for resolving the CURIEs
- Data **interoperability frameworks** like e.g. Biomedical Research Integrated Domain Group (**BRIDG**)

Metadata standards:

- Dublin Core Metadata (**DC**): set of 15 basic metadata elements
- Data Catalog Vocabulary (**DCAT):** interoperability between catalogues
- MetaData Registry (**MDR, ISO/IEC 11179:2023)**: how to maintain database of metadata
- Open Archives Initiative Protocol for Metadata Harvesting (**OAI-PMH**): set of 6 HTML verbs

Use of ontologies:

- Systems Biology Ontology (**SBO**)
- Terminology for Description of Dynamics (**Teddy**)
- Kinetic Simulation Algorithm Ontology (**KiSAO**)
- Provenance, Authoring, and Versioning (**PAV**)

Implementation guide – Metadata annotation

Metadata annotation:

- Triplet phrases: ***subject – predicate – object***"
- the recommended predicates are mostly '**is**' or '**isVersionOf**' (see Table 2 in **ISO 20691** for other predicates)
- done by embedding **RDF <annotation> elements** into XML-based data files

Example:

```
<rdf:Description rdf:about="./MyModel.xml#meta2">
    <dcterms:description>Cardiomyocyte cytosolic ATP concentration</dcterms:description>
</rdf:Description>
```

Checks for data quality:

- Completeness (missing values)
- Plausibility (range checks, cross-reference checks)
- domain-specific quality formats, for instance .**mzQC** for proteomics; in genomics / sequencing the quality information is often part of the data file (**FASTQ** …)

Implementation guide – Executing models

Model parameterization:

- In SBML: ***'Parameter'*** and ***'Constraint'*** components
- SED-ML: **Parameter** class
- Other: tabular parameter estimation (**PETab**) format encoding the model parameter information

Model execution:

- readers for **SED-ML**, unzipper for **OMEX** , … must be present in the execution environment

Depending on the model type, a proper **model solver** for running the simulation must be available

- for **deterministic** SBML models: **RoadRunner**, **CellDesigner**, **Copasi**, **Morpheus**, the **SBMLToolbox** or the **systems biology simulation core algorithm**
- **Gillespie algorithm** for **stochastic** simulations, discrete-event simulations (**DES**) and multi-agent-based simulations (**MABS**).

Implementation guide – Executing environment

Execution environment:

- Workstation
- High-Performance Cluster (**HPC**)
- cloud (Amazon **AWS**, **Google Cloud** or **MS Azure**)

→a **Docker** resp. **Apptainer** container should be available

Runtime environments:

- C (libc, msvcrt.dll)
- CLR (Common Language Runtime)
- JRE (Java Runtime Environment)
- Julia, Jupyter, Mathematica, Matlab / GNU Octave, Python, R, …

Workflow execution engine:

- Support for Common Workflow Language **(.cwl)** files
- Support for workload manager Simple Linux Utility for Resource Management (**Slurm**)
- Examples are **Arvados**, **Toil**, **StreamFlow**, **Sapporo** and **yadage**

Implementation guide – Verification and validation

Proposal is to follow the procedure described by the **ASME V&V 40** standard, see the Toward Good Simulation Practice (**GSP**) book.

Table 1: Terminology used by the ASME V&V 40 standard for quality assessment

Quality Term	Description	Evidence Type
Verification	Did you solve the underlying mathematical model correctly?	Mathematical Evidence
Validation	Does the underlying mathematical model correctly represent the reality of interest?	Experimental Evidence
Uncertainty Quantification	What is the uncertainty in the inputs (e.g., parameters, initial conditions), and what uncertainty in the outputs results from that?	Statistical Evidence
Applicability	How relevant is the validation evidence to support using the model in the context of use?	Engineering Judgement
Credibility	Based on the available evidence, is there trust in the predictive capability of the computational model for the context of use (CoU)?	Engineering Judgement

Implementation guide – Verification and validation

- Definition of the scientific / medical **question of interest (QoI)** and the **context of use (CoU):** The CoU is a complete description of the planned modelling use and defines the role and scope of the model used to address the question of interest
- **model risk** (with its two components **model influence** and **decision consequence**) - the possibility that the results of the model simulation are wrong and lead to negative consequences for the patient - must be assessed
- **applicability** of a model is given by the evidence to support the use of the model in the defined CoU
- **risk-informed credibility assessment** is performed which encompasses the **three credibility factors**
  - model **verification** with the two factors **code verification** (source code or algorithmic errors) and **calculation verification** (discretization or iterative errors)
  - model **validation** asks if the model can correctly simulate reality, e.g., the correctness of the underlying model assumptions and approximations
  - **uncertainty quantification** (**UQ**): sensitivity analysis to determine how sensitive the model output reacts to uncertainties in the model assumptions and input parameters

Implementation guide – Reporting, visualization and archiving

- **Reporting:**
  - SBML models: simulation results should be stored in Systems Biology Results Markup Language (**SBMRL**) files.
  - Consolidated Standards of Reporting Trials (**CONSORT**) for clinical trials
  - Strengthening the Reporting of Observational Studies in epidemiology (**STROBE**)
  - Standards for the Reporting of Diagnostic accuracy studies (**STARD**)
  - Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (**TRIPOD**)

Analogous set of regulations for Artificial Intelligence reporting (suffix **–AI**)

- **Visualization:** visual representations of signaling, metabolic and gene regulatory pathways
  - Systems Biology Graphical Notation (**SBGN**)
  - Biological Pathway Exchange (**BioPAX**)
  - …
  - **Archiving:** all the data, models, parameters, and simulation results must be archived according to the General Data Protection Regulation (timeframe, consent, …)

Implementation guide – EHR data and CDS

Electronic Health Record (EHR) data:

- should follow the European Electronic Health Record Exchange Format (**EEHRxF**) recommendations; can be implemented as
  - HL7 Fast Healthcare Interoperability Resources (**FHIR**) profiles
  - Integrating Healthcare Enterprise (**IHE**) profiles
  - Open Electronic Health Record (**OpenEHR**) format
  - For integrating EHR data with research data, the Biomedical Research Integrated Domain Group (BRIDG) standard can be used.

Clinical decision support (CDS) systems:

- knowledge representation the **Arden syntax** 3.0
- encoding of clinical decision support logic by
  - Clinical Quality Language (**CQL**) using **clinical quality measures** (**CQM**): measures performance on population health in response to the delivery of health care services
  - Clinical Decision Support Hooks **(CDS Hooks)**
  - Substitutable Medical Applications and Reusable Technology for CDS (**SMART on FHIR**)

Implementation guide – Building an approved model

- The general **model building process** should be done according to the ISO/TS 9491-1:2023 ("Predictive computational models in personalized medicine research – Part 1: Constructing, verifying and validating models") standard respecting the model formatting rules described in ISO 20691:2022 ("Requirements for data formatting and description in the life sciences").
- Iterating the **model execution cycle** consisting of the steps
  - model building/adaption
  - parametrization
  - execution
  - validation, verification and uncertainty quantification (**VVUQ**) → credibility

  until the model credibility is high enough to get regulatory approval at an official **Health Technology Assessment (HTA)** admission office, e.g. **FDA** or **EMEA**.
- Execute the approved model on the targeted execution environment. Before execution, the used patient data must be prepared according to the description in *section "Data handling"*.

The END – Questions, Comments, Remarks, Ideas, …

EDITH - European Virtual Human Twin
http://www.edith-csa.eu

Deliverables available under the tab 'Dissemination/Materials':
https://www.edith-csa.eu/materials/

Indication of interest via the contact form on site
https://www.edith-csa.eu/contact

EDITH

EDITH project has received funding from the EU H2020 Research and Innovation Programme, under Grant Agreement n. 101083771

Question to the audience: Implementability

Do you see any problems for the **implementability** of the presented standards on the **EDITH simulation environment**, especially with respect to the

- Catalogue / Repository
- Simulation platform
- Workflow execution engines

Questions to the audience: Missing standards, terminologies or services

- How can SDOs and standardization initiatives help to build a standard based VHT infrastructure?
- Are there relevant standards existing, but missing in our standards document?
- Are there domains for which proper standards or terminologies are still not defined?
- Is there a need for basic services, e.g., for a dedicated terminology service for VHTs?

Question to the audience:

The COMBINE consortium plans to define a standard for the semantic description for digital twins of living systems.

Which metadata do we need to semantically describe a life-science DT?

Biological metadata:

- Organism (human) → Virtual Human Twin (**VHT**)
- Purpose (diagnostic, prognostic, therapy choice, therapy optimization)
- System level (molecular (pathways), subcellular (organelles), cell, tissue (intercellular), organ, organ system, whole body, communities (epidemiological or population statistical models)
- Integration into multiscale models (cell-cell signaling (tissue factors), nervous signals, hormones, microbiome interactions (skin, gut, …), PBPK (physiology-based pharmacokinetic model)
- Model types (physiological / pathophysiological, biomechanical, imaging, systems biology / systems medicine / systems pharmacology (PK / PD) / systems toxicology)

Question to the audience:

The COMBINE consortium plans to define a standard for the semantic description for digital twins of living systems.

Which metadata do we need to semantically describe a life-science DT?

Biological metadata (continued):

Question to the audience:

The COMBINE consortium plans to define a standard for the semantic description for digital twins of living systems.

Which metadata do we need to semantically describe a life-science DT?

Technical metadata:

- Components of a DT (data, model or ensemble of models, comparator: divergence between the predicted state and observed state, actuator: does something in response to comparator output, control: decision what to do, i.e. model reacts to events)
- Input + Output data (data type, data format, frequency of data acquisition, availability, reliability, security, units)
- Integration into multi-scale models (VHT as a collection of DTHs: modularization --> need for interfaces)
- Model validation (VVUQ, identifiability, observability, reproducibility, explainability; Problem for personalized models: validation can only be done retrospectively)
- Parametric variability (How to adapt a generic model to become an individual, i.e. personalized DT ?)

Question to the audience:

The COMBINE consortium plans to define a standard for the semantic description for digital twins of living systems.

Which metadata do we need to semantically describe a life-science DT?

Technical metadata (continued):

- Language runtime (C (libc, msvcrt.dll), CLR, JRE, Julia, Jupyter, Mathematica, Matlab / GNU Octave, Python, R, …)
- Execution hardware (workstation, HPC cluster, cloud)
- Execution environment (Containerized, WfExS – Workflow Execution Service, Web application (running in the browser, …)
- Container type (Docker, Apptainer)
- Execution / solver software (COPASI, libroadrunner, Morpheus, Tellurium, Vivarium, XPPAUT)
- Model type (model type (ABM, algebraic equations, Bayesian, Boolean, Graphical, linear equations, metabolic network, ODE, PDE, Petri, stoichiometric, …)
- Model format (CellML, FieldML, SBML, R, XPP, …)

Question to the audience:

The COMBINE consortium plans to define a standard for the semantic description for digital twins of living systems.

Which metadata do we need to semantically describe a life-science DT?

Technical metadata (continued):

- Data format (XML, RDF, JSON, JSON-LD, YAML, FHIR, Apache Avro, protobuf, …) for patient / healthcare data files

Question to the audience:

The COMBINE consortium plans to define a standard for the semantic description for digital twins of living systems.

Which metadata do we need to semantically describe a life-science DT?

Legal metadata:

- Regulatory aspects (describing clinical studies; using standardized medical terminologies e.g. SNOMED-CT, Logical Observation Identifiers Names and Codes (LOINC), RxNorm, …)
- Privacy / data protection issues (anonymization, consent, **GDPR**)
- Data security
- Intellectual Property Rights (**IPR**)
- License information (list from the Open Software Initiative (**OSI**) or Software Package Data Exchange (**SPDX**)), e.g. Creative Commons Attribution 4.0, …