EDITH

Building the European Virtual Human Twin

Call: Accelerating best use of technologies (DIGITAL-2021-DEPLOY-01)
Work program year: DIGITAL-2021-2022
Topic: ID DIGITAL-2021-DEPLOY-01-TWINS-HEALTH
Grant Agreement No: 101083771

Task 4.2
EDITH standards implementation guide (IG)

Actual date of submission to Zenodo: 17 January 2024
Start of the project: 01 October 2022
End date: 30 September 2024

EDITH is a Coordination and Support Action funded by the Digital Europe program of the European Commission under the grant agreement n. 101083771.

Reference

Name	EDITH_T4.2: EDITH standards implementation guide (IG)
Lead beneficiary	HITS
Author(s)	Gerhard Mayer (HITS), Martin Golebiewski (HITS)
Dissemination level	Public
Type	Report

Version log

Date	Version	Involved	Comments
17/01/2024	V1.0	Gerhard Mayer (HITS), Martin Golebiewski (HITS)	First draft, final version for Paris Meeting submitted to Zenodo

Executive summary

The implementation guide is the Task 4.2 of the Coordination and Support Action (CSA) of the EDITH (Ecosystem Digital Twins in Healthcare) project. It is a shortened and, in some points more concrete version of the much longer standards document. The aim of the implementation guide is two-fold: First it gives hints to the modelers, which steps they should follow in the model building process and which standards, terminologies, and guidelines (depending on their modelling domain) they should use in defining their biomedical and healthcare models. Second it is intended as a practical guide for implementers giving hints, which standards, terminologies, and guidelines should be supported in the long-term by the simulation environment consisting of the repository, the simulation platform, and the workflow execution engines. Initially it suffices if they support all formats and annotations used by the demonstrator use cases.

Remarks and comments on how to improve the implementation guide are very welcome.

T4.2: EDITH standards Implementation Guide (IG)

Table of contents

1. Introduction

The computational modelling & simulation should be in accordance with the best practices of the Toward Good Simulation Practice (**GSP**) guidelines of the Avicenna Alliance and the Modeling Good Research Practices [1] defined by the International Society for Pharmacoeconomics and Outcomes Research (**ISPOR**) for Health Economics and Outcomes Research (**HEOR**), and the Society for Medical Decision Making (**SMDM**).

This document provides a guideline for using and implementing standards, terminologies, and metadata guidelines (listed in detail in the annex of the EDITH-T2.3 document “Analysing current landscape of standards, identifying needs and gaps") when setting up, executing and archiving virtual human twins. To get an overview and access information on such standards, terminologies, and metadata guidelines there is the EDITH Fairsharing collection available.

In the European context, for the use of standard formats there is the general rule - especially in areas, where data protection laws are affected - that whenever a European standard is available, it should preferably be used, e.g., use of the European Electronic Health Record Exchange Format (**EEHRxF**) for exchanging EHR data.
For the artificial intelligence models at least the following 3 standards mentioned in the **European AI Act** shall be followed [2]:

- IEEE P7003 Standard for Algorithmic Bias Considerations [3]
- IEEE P7001/D4 Draft Standard for Transparency of Autonomous Systems [4]
- IEEE 7000 Standard Model Process for Addressing Ethical Concerns during System Design

In the following an overview about the most important points for implementing the use of standards and terminologies is given.
It is emphasized here that both the models and data as well as the EDITH infrastructure components catalog/repository and the simulation platform should adhere as far as possible to established (official or alternatively community) standards for computational modelling & simulation in the areas of systems biology, systems medicine, systems pharmacology, and systems physiology.
This conformance to standards means to adopt and promote generally applicable and discipline specific operating procedures, guidelines, regulations, formats, and terminologies [5].

2. Data handling

2.1 Data preparation

The data for construction, validation, execution, description and archiving of virtual human twin components shall be prepared by cleansing and formatting it into accepted standard formats, in accordance with **ISO 20691:2022** and **ISO/TS 9491-1:2023**. The data preparation comprises the following steps:

- sampling the data
- data formatting and harmonization, e.g., lab value concentrations shall 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 shall 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.
- data integration, either on the personal or on the variable level.
- adding data provenance information.
- defining who can access the data.

This procedure ensures reproducibility, interoperability, data completeness and high data quality. After data preparation, the data should be **FAIR** (Findable, Accessible, Interoperable, Reusable). In addition, provenance information according to the **ISO 23494 series** shall be added to the data. Such FAIR data shall possess a unique identifier, be linkable, have an assigned license, and be enriched/annotated with metadata that describes various attributes, such as the disease, tissue, cell type, and modelling parameters.

For the input and output interfaces of a virtual human twin one should avoid using spreadsheets (.csv) or text files without having information clearly describing the meaning of the cells. Instead of that, the usage of file formats with a schema or serialization description describing the contents should be used, e.g., **ObjTables** for .csv files. Instead of .txt files **.json** files with a defining **JSON schema** or **.xml** files with a **XMLSchema Definition (.xsd)** should be used. For binary files a standardized serialization format like **Avro Apache** or **protobuf** is strongly recommended.

2.2 Data integration

For data integration and interoperability, a standardized method for the unique identification of all biochemical entities shall be used (see also **ISO 20691**), e.g. by applying persistent identifiers like compact Uniform Resource Identifiers (**CURIEs**). These identifiers can be resolved to HTML links by using the **Bioregistry** repository, which standardizes the usage of CURIEs. Such CURIEs have the general form "**prefix:local unique identifier**", where the prefix encodes the resource.

2.3 Annotation with metadata

For basic metadata, the use of standards like Dublin Core Metadata (**DC**), Data Catalog Vocabulary (**DCAT)** or MetaData Registry (**MDR, ISO/IEC 11179:2023)** shall be used. After checking which are the proper ontologies and the minimum required information for the considered data and modelling domain, one should annotate the domain specific data with terms from these ontologies. For the exchange of the metadata a standard protocol shall be used, e.g. Open Archives Initiative Protocol for Metadata Harvesting (**OAI-PMH**).

For the semantic annotation of data and models terms from established domain-specific terminologies shall be used, e.g. from SNOMED-CT, UMLS, or similar for medical data, or from Systems Biology Ontology (**SBO**), Terminology for Description of Dynamics (**Teddy**), Kinetic Simulation Algorithm Ontology (**KiSAO**), or similar for systems biology / medicine models. Depending on the modeling domain also terms from other domain-specific ontologies can be used for annotation.

Recommendations on how to use semantic annotations for computational models are given by [6]. For instance, the syntax of the semantic annotations are triplet phrases of the form "***subject* – *predicate* – *object***". The recommended predicates are mostly 'is' or '**isVersionOf**', but other predicates are possible as well, see Table 1 and 2 of the **ISO 20691** standard documents.

If an XML-based format is used, the annotation itself should be done by embedding **RDF <annotation> elements** into the XML-based data files [6]. Also other formats like JSON can be used (preferably also as an RDF annotation), if the annotation is compatible with the W3C “Linked Open Data” concept.

2.4 Check of data quality, plausibility, and completeness

Missing values, correct data types and the presence of the corresponding units shall be checked. If sensible and acceptable, one can take data imputation methods for missing values into account. Otherwise, the datasets with missing values shall be removed before the model building to avoid biased conclusions. Then the **plausibility** of the data shall be checked (e.g. range checks, cross-reference checks). If there are domain-specific quality formats available, they should be used, e.g., **mzQC** (previously named **qcML**) for quantitative proteomics data. Further quality validation recommendations are defined in **ISO/TS 9491-1**.

3. Executing models on patient/healthcare data

3.1 Model parameterization

The model parameters like initial values, boundary conditions, and constraints shall be specified. In SBML models these are defined in the ***‘Parameter’*** and ***‘Constraint’*** components in the SBML model files. For other systems biology models one can consider using the tabular parameter estimation (**PETab**) format encoding the model parameter information. Constant parameters can also be defined in the Parameter class of SED-ML.
Best practices for computational modelling & simulation are described in the Toward Good Simulation Practice (**GSP**) book.

3.2 Model execution

3.2.1 Model description

Simulation Experiment Description Markup Language (**SED-ML**) should be used to describe the simulation experiment setup, if feasible. It can be used to describe the data, the models, the simulation setup, the simulation procedures and how the results of the simulations look like. Therefore, readers for **SED-ML**, **SBML**, **OMEX** and other model file formats should be present in the execution environment if used to describe the simulation setup and procedure. For AI/ML models the model execution puts the trained model to work on real input data.

3.2.2 Model solver

Depending on the model type, an appropriate model solver for running the simulation should be used. Examples are **RoadRunner** [7], **CellDesigner** [8], **Copasi** [9], **Morpheus** [10] and the **SBMLToolbox** [11] for **deterministic** SBML models. An alternative is the **systems biology simulation core algorithm** [12], a library of different numerical solvers for numerical integration of the set of differential equation systems defined in an SBML file.

For **stochastic** simulations the **Gillespie algorithm** [13-15] is the method of choice. In an extended form it can also be used for the simulation of discrete-event simulations (**DES**) and multi-agent-based simulations (**MABS**) [16].

3.2.3 Targeted execution environment

Next, choose the execution environment on which the model should run. This can either be a workstation, a High-Performance Cluster (**HPC**) or in the cloud (e.g. Amazon **AWS**, **Google Cloud** or **MS Azure**). For that a **Docker** resp. **Apptainer** container should be available. It allows one to execute the model in the chosen environment. For execution in an HPC environment, a Docker container shall be converted first into an Apptainer container.

In addition to the model solvers, the required **runtime environments** (e.g. C (libc, msvcrt.dll), CLR (Common Language Runtime), JRE (Java Runtime Environment), Julia, Jupyter, Mathematica, Matlab / GNU Octave, Python, R, …) shall be available in the execution environment.

3.2.4 Execution as workflow

For execution on a workflow execution engine the steps of the model execution should be formulated in a Common Workflow Language **(.cwl)** file. It's recommended that a workflow execution engine, which supports the workload manager Simple Linux Utility for Resource Management (**Slurm**), is used.

Therefore **Arvados**, **Toil**, **StreamFlow, Sapporo** [17] and **yadage** are considered eligible.

3.3 Validation and verification of modelling results

An important standard is **ASME V&V 40** [18], originally defined by the American Society of Mechanical Engineers for risk-based assessment of the quality and model credibility of medical devices. The model quality assessment starts with a clear 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. In the next step the **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 - shall be assessed. The **applicability** of a model is given by the evidence to support the use of the model in the defined CoU, considering the risk. Applicability therefore depends on the closeness of the model input parameters values to their values of the CoU. For instance, one can ask if the range of model input parameters covers the whole parameter space of the CoU. Then a **risk-informed credibility assessment**, which encompasses the three credibility factors model **verification**, model **validation**, and **uncertainty quantification** (**VVUQ**), is performed (see *Table 1*). Uncertainty quantification is a form of sensitivity analysis to determine how sensitive the model output reacts to uncertainties in the model assumptions and input parameters. The model **verification** consists of the two factors code verification (source code or algorithmic errors) and calculation verification (discretization or iterative errors). The model **validation** asks if the model can correctly simulate reality, e.g., the correctness of the underlying model assumptions and approximations. This can be done by comparing the model forecast with measured experimental values.

The use of ASME V&V 40 for the evaluation of new drugs from *in silico* trials is currently discussed by regulatory authorities like FDA and EMA [19].

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 is the resultant uncertainty in the outputs?	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?	Engineering Judgement

3.4 Reporting and visualization of modelling results

The parameters and simulation results of SBML models should be stored in Systems Biology Results Markup Language (**SBMRL**) files. The modularization of hierarchical SBML models can be described using the SBML Level 3 package for hierarchical model composition (**SBML-comp**) [20]. Modularized models allow bridge processes at the cellular level with processes on the tissue level to construct multiscale organ models [21].

For domain-specific reporting the corresponding guidelines should be followed. Examples are Consolidated Standards of Reporting Trials (**CONSORT**) for clinical trials and Strengthening the Reporting of Observational Studies in Epidemiology (**STROBE**) for epidemiological studies. The Agency for Healthcare Research and Quality (**AHRQ**) defined rules for the reporting of modeling and simulation studies for Health Technology Assessment (**HTA**).

The visualization of systems biology and systems medicine data can be done by using disease maps, which are a mapping of gene activities and/or protein or metabolite concentrations onto visual representations of signaling, metabolic and gene regulatory pathways, stored in standard formats like Systems Biology Graphical Notation (**SBGN**) or Biological Pathway Exchange (**BioPAX**). By proper color-coding one can visualize concentration changes of the relevant biomolecules over time and distributed over the whole network.

3.5 Archiving

According to the General Data Protection Regulation (**GDPR**) all the data, models, parameters, and simulation results derived from clinical data shall be recorded and archived for up to 30 years (either in the EDITH infrastructure or in an appropriate external repository), so that the whole simulation can be reproduced later.

3.6 Electronic Health Record data

Whenever possible, routine Electronic Health Record (**EHR**) data should use the European Electronic Health Record Exchange Format (**EEHRxF**). It can be implemented using either the HL7 Fast Healthcare Interoperability Resources (**FHIR**) profiles, Integrating Healthcare Enterprise (**IHE**) profiles or alternatively the Open Electronic Health Record (**OpenEHR**) format.

In case these EHR data shall be made interoperable with clinical research and/or observational epidemiological health data encoded in Observational Medical Outcomes Partnership (**OMOP**) format, the Biomedical Research Integrated Domain Group (**BRIDG**) [22] standard can be used.

3.7 Clinical decision support systems (CDSs)

For knowledge representation the **Arden syntax** [23], and for encoding of clinical decision support logic one of the following should be used [24]:

- Clinical Quality Language (**CQL**)
- Clinical Decision Support Hooks **(CDS Hooks)**
- Substitutable Medical Applications and Reusable Technology for CDS (**SMART on FHIR**)

4. Building an approved model

4.1 Model building

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").

Then, after sensible parameterization build a sensible, usable, and credible model according to the rules of the Toward Good Simulation Practice (**GSP**) book, by iterating the model execution cycle described in *section 3 ("Executing models on patient/health data")*. The cycle consists of the steps model building/adaption, parametrization, execution, validation/verification, and uncertainty quantification (see *section 3.3 "Validation and verification of modelling results"*) and is repeated until the model credibility is high enough to get regulatory approval.

4.2 Getting regulatory approval

Get regulatory approval at an official **Health Technology Assessment (HTA)** admission office, e.g., Food and drug Administration (**FDA**) or European Medicines Agency (**EMEA**). For getting approval the model shall be highly credible, assessed by the credibility assessment procedure for *in silico* models as described in [19] and in *section 3.3* "Validation and verification of modelling results".

4.3 Execute the approved model

Execute the approved model on the targeted execution environment (see *section 3 "Executing models on patient/health data")*. Before execution, the data shall be prepared according to the description in *section 1 ("Data handling")*.

Acknowledgement

EDITH is a Coordination and Support Action funded by the European Commission under the Digital Europe Programme (DIGITAL) under grant agreement number 101083771.

References

1. Caro JJ, Briggs AH, Siebert U, Kuntz KM, Force I-SMGRPT: **Modeling good research practices--overview: a report of the ISPOR-SMDM Modeling Good Research Practices Task Force-1**. *Med Decis Making* 2012, **32**(5):667-677.
2. **AI Watch: Artificial Intelligence Standardisation Landscape Update**. In. Luxembourg: European Union; 2023: 41.
3. Koene A, Dowthwaite L, Seth S: **IEEE P7003™ standard for algorithmic bias considerations**. In: *Proceedings of the International Workshop on Software Fairness*. 2018: 38-41.
4. Winfield AFT, Booth S, Dennis LA, Egawa T, Hastie H, Jacobs N, Muttram RI, Olszewska JI, Rajabiyazdi F, Theodorou A *et al*: **IEEE P7001: A Proposed Standard on Transparency**. *Front Robot AI* 2021, **8**:665729.
5. Erdemir A, Mulugeta L, Ku JP, Drach A, Horner M, Morrison TM, Peng GCY, Vadigepalli R, Lytton WW, Myers JG, Jr.: **Credible practice of modeling and simulation in healthcare: ten rules from a multidisciplinary perspective**. *J Transl Med* 2020, **18**(1):369.
6. Neal ML, Konig M, Nickerson D, Misirli G, Kalbasi R, Drager A, Atalag K, Chelliah V, Cooling MT, Cook DL *et al*: **Harmonizing semantic annotations for computational models in biology**. *Brief Bioinform* 2019, **20**(2):540-550.
7. Welsh C, Xu J, Smith L, Konig M, Choi K, Sauro HM: **libRoadRunner 2.0: a high performance SBML simulation and analysis library**. *Bioinformatics* 2023, **39**(1).
8. Matsuoka Y, Funahashi A, Ghosh S, Kitano H: **Modeling and simulation using CellDesigner**. *Methods Mol Biol* 2014, **1164**:121-145.
9. Bergmann FT, Hoops S, Klahn B, Kummer U, Mendes P, Pahle J, Sahle S: **COPASI and its applications in biotechnology**. *J Biotechnol* 2017, **261**:215-220.
10. Starruss J, de Back W, Brusch L, Deutsch A: **Morpheus: a user-friendly modeling environment for multiscale and multicellular systems biology**. *Bioinformatics* 2014, **30**(9):1331-1332.
11. Keating SM, Bornstein BJ, Finney A, Hucka M: **SBMLToolbox: an SBML toolbox for MATLAB users**. *Bioinformatics* 2006, **22**(10):1275-1277.
12. Keller R, Dorr A, Tabira A, Funahashi A, Ziller MJ, Adams R, Rodriguez N, Le Novere N, Hiroi N, Planatscher H *et al*: **The systems biology simulation core algorithm**. *BMC Syst Biol* 2013, **7**:55.
13. Gillespie DT: **Exact stochastic simulation of coupled chemical reactions**. *The Journal of Physical Chemistry* 2002, **81**(25):2340-2361.
14. Gillespie DT: **Stochastic simulation of chemical kinetics**. *Annu Rev Phys Chem* 2007, **58**:35-55.

15. Cai X: **Exact stochastic simulation of coupled chemical reactions with delays**. *J Chem Phys* 2007, **126**(12):124108.

16. Montagna S, Omicini A, Pianini D: **Extending the Gillespie's Stochastic Simulation Algorithm for Integrating Discrete-Event and Multi-Agent Based Simulation**. In: *MABS 2015*. Springer 2016.

17. Suetake H, Tanjo T, Ishii M, P. Kinoshita B, Fujino T, Hachiya T, Kodama Y, Fujisawa T, Ogasawara O, Shimizu A *et al*: **Sapporo: A workflow execution service that encourages the reuse of workflows in various languages in bioinformatics**. *F1000Research* 2022, **11**.

18. Viceconti M, Emili L: **Toward Good Simulation Practice**: Springer; 2024.

19. Viceconti M, Pappalardo F, Rodriguez B, Horner M, Bischoff J, Musuamba Tshinanu F: **In silico trials: Verification, validation and uncertainty quantification of predictive models used in the regulatory evaluation of biomedical products**. *Methods* 2021, **185**:120-127.

20. Smith LP, Hucka M, Hoops S, Finney A, Ginkel M, Myers CJ, Moraru I, Liebermeister W: **SBML Level 3 package: Hierarchical Model Composition, Version 1 Release 3**. *J Integr Bioinform* 2015, **12**(2):268.

21. Shahidi N, Pan M, Safaei S, Tran K, Crampin EJ, Nickerson DP: **Hierarchical semantic composition of biosimulation models using bond graphs**. *PLoS Comput Biol* 2021, **17**(5):e1008859.

22. Becnel LB, Hastak S, Ver Hoef W, Milius RP, Slack M, Wold D, Glickman ML, Brodsky B, Jaffe C, Kush R, Helton E: **BRIDG: a domain information model for translational and clinical protocol-driven research**. *J Am Med Inform Assoc* 2017, **24**(5):882-890.

23. Csarmann A, Zeckl J, Haug P, Jenders RA, Rappelsberger A, Adlassnig KP: **Arden Syntax on FHIR**. *Stud Health Technol Inform* 2023, **305**:423-424.

24. Taber P, Radloff C, Del Fiol G, Staes C, Kawamoto K: **New Standards for Clinical Decision Support: A Survey of The State of Implementation**. *Yearb Med Inform* 2021, **30**(1):159-171.