Cytokine semantic module model

Files

Please, review the model in the following order:

- PNG: The first model is the ejp core model with the cytokine view, I.e. it is the core model with added semantics to meet the cytokines interpretation requirements. The second model is the complete semantic model for the cytokines. The third model envisions futurible extensions that can be added once we have clinical data or other related data.
    1. ejp_core_model_cytokines_view.png
    2. cytokines_semantic_model.png
    3. cytokines_extended_semantic_model.png
- OWL: Domain and Range have to be added. Also Descriptions, ontology URIs, polishing...
- Validation: to be done (maybe ShEx ?)

Description

The model is based on several existing models and raw data:

- **Qinqin cytokine** model
- **ISMB** model: focused only on describing a quantitative trait using COVID as use case https://github.com/NuriaQueralt/BioHackathon/tree/master/bh20-ontology-qt
- **SIO measurements** model: Model for measurement processes in science. This is the base for the the EJP-ERN-CDE-Core model https://github.com/MaastrichtU-IDS/semanticscience/wiki/DP-Measurements
- **EJP-ERN-CDE-Core** model: Process-based core model for ERN. This is the base for our model and it is kind of the most complete model that overlaps the previous ones https://github.com/ejp-rd-vp/ERN-common-data-elements/wiki/Core-model-SIO
- The link with **Cesar patient** and **Qinqin covid-hierarchy** semantic modules

Revision 12 June 2020

Cesar notes

Hey Nuria, I've analyzed the models and I think they all are very good. Just wrote some few considerations here, but just some thoughts... Let me know what you think. ‘Model.docx’

Qinqin notes

The models themselves are very good, (except for the arrow direction between ward and hospital, which I guess it should be other way around). It is just a bit unclear for me about the correspondences between the models and OWLs, e.g. which property in OWL corresponds to the "a" properties shown on the graph? Another thing is that the OWL include classes/terms (e.g. "Age", "Device", "Time") that have been defined in some standard ontologies, therefore it will be good to use the existing terms from standard ontologies.

Rajaram notes

The models looks really rich and good. But why sample has different relationship in measurement process? why not :used relation to link sample and measurement process? What I am trying to see is if the "sampling process" generate multiple output(s). I am not sure if it does but that happens then we need precise relationship between "measurment-process" and "sampling-process" right? then ur "measurment-process" points to one of the output of the "sampling-process" right? I mean the exact sample used in the "measurment-process" Núria: Good question! Atm we only have info from the sampling process, so I cannot answer if the sampling has only 1 or N outputs.

TC Minutes 12 June 2020

Participants: Cesar, Qinqin, Rajaram, Núria

Agenda: Discuss the model to be uploaded into Neo4j

Summary:

- Neo4j is public -> modify patient sensitive data (pick 10 rows and modify values randomly)
- Validate using user-friendly models and queries with LUMC experts: Small and FAIR (watch presentation from Marco)
- Email asking for a meeting with Jacqueline/Simone: we have to bring the value of interoperable and FAIR data. **Goal: Jacqueline prepare better XLS, more FAIR to populate OPAL warehouse.**
- Our FAIR work will be: Convert data from MONGO DB (backend of OPAL Warehouse), using RML or Cesar suggested software, into FAIR machine readable format such as RDF
- **Ask Jacqueline: Clinical_id, beat_id ???** (prov:hadPrimarySource): modeling as <measurement_process>- prov:hadPrimarySource -<castor DB>-dct:identifier-<clinical_id> take message home for Jacqueline: we don't know and we make mistakes modeling because data is not FAIR
- Rajaram: standard for the literal shape, how to link the sample with the measurement process (not urgent)
- Next step: upload few raws to HG setup
- Some useful links (suggested during the meeting):
    - RML Transformation Language https://rml.io/
    - Castor https://www.castoredc.com/
    - OPAL https://opal-llc.com/solutions/warehouse/
    - hadPrimarySource relation https://www.w3.org/TR/prov-o/#hadPrimarySource
    - wasInformedBy relation https://www.w3.org/TR/prov-o/#wasInformedBy
    - ShEx https://shex.io/
    - SHACL https://www.w3.org/TR/shacl/
    - Rajaram paper https://docs.google.com/document/d/10ZYhccZimpCGz7o5PPVZaOC4zlqIvuI-BejlZA9uZ4Q/edit

TC Minutes 16 June 2020

Participants: Jacqueline, Cesar, Qinqin, Núria

Agenda: Discuss the metadata fields from CASTOR DB with Jacqueline (Expert on parasites and data management). Jacqueline is on charge to add all different lab data measurements (done by different labs) into CASTOR DB, and then these data will be integrated with clinical data into the OPAL Warehouse.

Summary:

Núria: According to Jacqueline, the two most important fields to track are 1) measurement date, 2) patient. Also she said that the majority of the metadata is described in the dictionary PDF (she will upload the newest version because the one that is now in the sharepoint is obsolete)

record id: number for the collection of samples. For each sample, there is a ".1, .2, .3" for each sample material. Linked to a patient and to data sampling. Núria: identifier for all the samples collected from one patient on one date_sampling. It is the collection of samples of one patient on one date (sampling_date). If the sampling has been ordered, then one sampling process will have from one to multiple samples, in other words the relation (sampling)--(sample_output) is 1sampling --1..N samples (samples, i.e. blood tubes).

Order:

BEATAD: Beat Admission Order

BEATMO: Monday Order

BEATWO: Wednesday Order

Jacqueline thinks that describing the order is not important. Núria: Different values. Each value means a different set of samples/measurements ordered, e.g. BEATAD means only serum samples that implicitly means only cytokines will be measured. Depends on the order --> different blood tubes, for example: serum, plasma, heparine... 1 order --> 1 date_sampling --> 1..N samples (or tubes) --> same record_id!!! (1 record_id --> 1..N samples).

- Record ID format: 201708863 (20 year, 17 week, 08863 patient id). Núria: record_id instance links all the other values instances such as ward, age, measurement_date, sampling_date..

- Institute Abbreviation: not relevant since now it has only "LUMC". Still, it can be used in the future since it identifies the health institution.

- Record Creation: the exact moment when data was included in castor. Núria: Jacqueline thinks it is not relevant, because it is a timestamp that CASTOR system adds by default.

- Order: not relevant. However, it could be interesting if we could have filters regarding the order ID. Núria: It can be a handy filter, for example to query what kind of samples a patient has been ordered because then we can infer what kind of measurements have been done for that patient (remember, each sample type is linked to a specific lab where samples are sent to perform a specific type of measurement).

- Ward: the importance here is to look for people that are in IC or not. IC wards start with "IC" letters, but MCB4 is also an IC, not sure about the others. Jacqueline will create a variable for "IC/Non-IC". Núria: It is relevant and necessary to know if the ward is an ICU or not, but from the current data we don't know. We need that Jacqueline adds this annotation.

- Age: the age and the birth date can be different. So it would be better if the age is linked to the measurement date. Núria: FILTER. We already modeled this correctly, I.e age value instance is linked to the measurement_process instance. The birth_date is not tracked in CASTOR because it is sensitive data, that is it could be used to identify the patient, so we have to protect patient privacy.

- beat id: the patient ID (anonymized). Núria: anonymized patient ID for CASTOR database.

- clinical id: record ID that patients have in the clinical database (where the clinical data its stored). the record ID that is used for the patient in the castor clinical database. The record ID in the clinical database for the patient. The anonymized ID for the clinical database. Núria: ABSOLUTELY NECESSARY to link lab data to clinical data. Record ID for the anonymized ID used for the patient in the clinical database.

- lum_date_mes: the data (time) that the cytokines were analyzed in luminex. Núria: the date the measurement was done (timestamp), very RELEVANT.

- Panel (test KIT), you measure multiple cytokines in a kit, and there are four kits. Panel is a protocol that you follow (like a recipe). Núria: here panel is used equivalently as kit, so kit = protocol = panel = assay followed in the lab to count the amount of analyte or whatever measurement type is performed.

- It is always more than one sample, depending on the order. (there is a relation with order and the record id (repeats))

Núria: Jacqueline agreed to give us access to CASTOR to analyze the raw data by ourselves.

Schema namespaces

Base : <>

rdf

rdfs

xsd

Classes

sfdf

Identification of Instances

sfdf