**SwissPedHealth Contribution to TheHyve Data Model**

**DCC Comment:** *We would be very interested to know more about this contribution and how it relates to the SPHN Schema*

**Project specific concepts developed. Contribution to TheHyve data model**

The SwissPedHealth lighthouse project group's contribution to TheHyve data model focused on defining and refining the use of omic data, with initial prioritization of whole genome sequence (WGS) data, within the lighthouse project. We detailed the process for managing WGS data, which involves:

* Outlining the journey of WGS data from acquisition by SMOC (H2030GC) to comprehensive analysis, incorporating best practices in quality control (e.g. GATK pipelines), monitoring (file and subject tracking, QC pipelines/fastp, etc), alignment (raw data indexes to BAM), variant calling (raw data to VCF generation), and detection (ACMGuru output).
* Emphasizing data traceability to ensure the precise tracking of raw data generation, sample processing, and the application of specific algorithms and technologies (software and SOP version control), enhancing the integrity and reproducibility of data within TheHyve model.
* Contributing to TheHyve model with description of omic data generation. This included processes such as library preparation, sequencing assay, and sequencing analysis (key metadata necessary during analysis QC and final reporting).
* Defined where existing semantic data models were useful to integrate or which required additional common biomedical vocabularies, to therefore support the FAIRification of omics data; discussions on cardinality and other requirements for concepts in data generation and reporting such as sequencing instrument, genetic variant, variant descriptor (2023.08.01-2023.08.11 prep for consensus statements).
* Reviewing and providing feedback on the “Genomics statements” document used as input to build and validate the model.
* Coordination with TheHyve during development of the project specific concepts, providing feedback to ensure TheHyve data model is suitable as a basis to directly inherit our project specific concepts from
* Workshops and meetings with TheHyve (acknowledged in [doi.org/10.20944/preprints202312.0373.v1](http://doi.org/10.20944/preprints202312.0373.v1))
* Meetings

Our specific contributions include defining the minimal necessary information for tracking and analysis, leading to the HIVE's incorporation of essential features like sequencing instrument details, SOPs, QC metrics, and metadata related to sequencing assays. Following this, our group introduced new concepts for post-analysis reporting, encompassing analysis results, statistics, and other omics-specific features, further enhancing the schema's capacity for detailed data representation and utility in research and clinical settings. The lighthouse project concepts were developed with the idea of expanding the model from TheHyve.

**Summary**

* TheHyve produced concepts which allow tracking from sample to raw omic data delivered by SMOC (the pre-analysis genomic dataset).
* Our group contributed by helping to define the exact minimal information that we require, both for tracking and during analysis.
* The TheHyve defined key features from our consensus statements [[link to doc](https://docs.google.com/document/d/1LzdZO_k0LFf7GJZF66C4UVrLu29gc9Z0nbiTmRp-OnI/edit#heading=h.lan8xhso062y)] such as:
  + Sequencing instrument, SOP name/description/version control, QC metrics, library preparation, sample identification, metadata associated with sequencing assay.
  + Examples in slide 5,6,11,12,13 [link to doc](https://docs.google.com/presentation/d/1kZgCOov8eymOhLz_OqtgUmFnOJJRXJUX/edit#slide=id.p1).
* Our group subsequently used and extended these concepts and followed up with new downstream concepts for post-analysis reporting with key features such as:
  + Analysis, results, statistics, and other omic-specific features.
  + Examples in slide 3,7 [link to doc](https://docs.google.com/presentation/d/1kZgCOov8eymOhLz_OqtgUmFnOJJRXJUX/edit#slide=id.p1).

**Meeting Dates**

2023.06.20 Workshop - Bern

* Concept design for omics (meta)data
* Personalized Health Informatics, SIB, DCC, TheHyve, SwissPedHealth
* Slides: [link to doc](https://docs.google.com/presentation/d/1bhyWYgwdg4OBWXzDEOL7Y2HvsFaUizJq/edit#slide=id.p1)

2023.08.01 Collaboration prep

* Omic metadata standards, concepts prioritization, metadata required for trust, creating general concepts, suitable for multi-type omics data, anonymised examples of data and metadata.
* Draft concepts drive [link to dir](https://drive.google.com/drive/u/1/folders/1xIX6quemEjTi3AM1NeTtH2mJ-fdgHD_Y)
* Draft data modeling notes [link to doc](https://docs.google.com/document/d/1OabO93bdt73mdpAF99HhqniHxFoau03oT9rIr9LZOcc/edit)
* Anonymised examples of data metadataomic\_data\_metadata.tvs [link to doc](https://docs.google.com/spreadsheets/d/1601Ij928OOhd5x6W19lWQay_Z9CnkpS2UDATgUlChHw/edit?usp=sharing)
  + - 1 DNA file metadata
    - 1 RNA file metadata
    - 1 DNA variant report data

2023.08.11 Design meeting - online

* TheHyve / SwissPedHealth data modeling interview
* Design notes: [link to doc](https://docs.google.com/document/d/1OabO93bdt73mdpAF99HhqniHxFoau03oT9rIr9LZOcc/edit)
* Consensus statements: [link to google doc](https://docs.google.com/document/d/11ADiekAzfi9IqxBtRgZtdkNgFNeGE7Oq5AXJ6fbGiSc/edit#heading=h.lan8xhso062y)

2023.09.04 Draft concepts review

* Testing Dataset2RDF

20230907 Revision with TheHyve

* Graph v0 20230907\_SPHN Genomics Extension - v1 [link to doc](https://drive.google.com/file/d/1VwZP5KEbnNp9vAIwSLAtrqdb33q86m_g/view)

20230928 Revision with TheHyve

* SPHN genomics statements v0.2 [link to doc](https://docs.google.com/document/d/1LzdZO_k0LFf7GJZF66C4UVrLu29gc9Z0nbiTmRp-OnI/edit#)
* SPHN genomic statements doc v0.1 [link to doc](https://docs.google.com/document/d/11ADiekAzfi9IqxBtRgZtdkNgFNeGE7Oq5AXJ6fbGiSc/edit#heading=h.lan8xhso062y)

2023.10.03 SPHN theHyve workshop - Bern

* Finalise concepts
* Slides: [link to doc](https://docs.google.com/presentation/d/1pXQs_1AWnYHQKg7T1oBK04i1f521hYJF/edit#slide=id.p1)
* Validation of concepts for omics (meta)data; SwissPedHealth, DCC, TheHyve: Incorporated feedback (Oct), Finalized concept proposals (Oct), Review of concept proposals (Oct), Incorporation into the SPHN Dataset for 2024 Release (Nov), Release candidate (late Nov), Release 2024 (Jan)

2023.11.xx Submitted the omics extension to cover non-genomics omics.

2023.10.31 Final version

2024.01.01 Planned release of the SPHN Dataset and RDF Schema

* Genomic concepts to be included
* Form the basis for representing omics (meta)data in the National Data Streams.