DTL workshop – October 2nd

Exec summary:

In this session, moderated by Morris Swertz (UMCG) and Peter-Bram ‘t Hoen (RUMC), we started designing a roadmap for large-scale -omics and cross-omics digital infrastructure. The objectives of such a roadmap would be fully In line with the NIH data commons initiative presented in the keynote talk by Titus Brown. This initiative should find a translation to the Dutch setting?

The needs for such a digital infrastructure were exemplified in pitches of three large projects: X-omics (Alain van Gool, RUMC), Unified for Metabolic Diseases (Clara van Karnebeek, AMC), European Joint Programme for Rare Diseases (Martina Kutmon, MUMC)

In subsequent breakout sessions with very active contributions from the participants, the following questions were answered: What are the needs / expectations for such a digital infrastructure? What is already there? What needs to be developed and who can contribute to that?

From the breakout sessions, we drew the following conclusions:

The most important needs that need to be prioritized are:

1. easy access to data and tools, with derived needs of federated authentication and authorization, interoperability of data and services, and rich metadata (including info on quality and level of curation, and tools to perform annotations).
2. incentives for providing FAIR data and tools. Example: a PhD student can substitute a traditional research chapter in his or her thesis by the provision of a dataset for reuse scoring high on the FAIRmetrics. Other example: institutions need to take into account reuse of data, tools and services when evaluating their employees (and not just the number of citations and/or h-factor).

DTL (together with HealthRI , ELIXIR etc) and we as a community can play an important role

1. to provide more insight in what is already there, in terms of projects, tools, services, experts and facilities
2. to lobby for use of funding for sustaining data, tools and services after the project
3. to create incentives and to lobby for adoption of valuing the provision of interoperable data, tools and services (by organisations, funders).

National agenda / roadmap for -omics infrastructures is basically about translating NIH data commons to the Dutch setting

Breakout Jaap Molenaar (moderator); Alida Kindt Dunjko (reporter)

- User-driven / use case driven
- From monolithic to federated solutions.

- Seamless access to data storage
- Interoperable data
- Same user accounts; easy application for access
- Reproducibility
- Flexible
- Workflows for returning questions
- Metadata (provenance); samples and measurements: Tooling for analysis and visualization should enforce this, think about benefits. Ontology-based annotation.
-

Current tools need to be adapted, made more interoperable, modified according to an interoperable guideline, associated with

How can we incentivize making investments in increasing interoperability of data and services. Start with practical reference implementations.

RISKTOXNET: developing services in the cloud

ELIXIR proteomics community:

Hyve: software layer for integration of data sources.

Elsevier: additional cloud services (like Mendeley): Olaf Lodbrok.

Business models. Do we want to pay for services?

DOI/GUID: continuity.

Role for funders: make sure that budget is set aside for me.

Elsevier: how can we contribute / facilitate

Varian: PHT/distributed learning

Registry in a box: Sander de Ridder. Connect to the data linkage plan?

Alida Kindt Dunjko: involve her in data analysis WP from the Hankemeier group.

Plenary feedback:

Mark Santcroos (LUMC): many things are already there. 100M not needed. Access to data and tools needs to be improved. Quality and reproducibility.

Already there: Research cloud; Radboud DRE; Data4LS; local expertise ; federated access and identity management; eduroam

Needs: Existing components asier to link and easier to combine, easier to link to

Funding agencies should avoid fragmentation of solutions.

Patients and citizens in control, navigator seat (not driver seat).

No reinvention of wheels.

Gerben Stouten (TUD): Access to data, standardization of data sources from different types of equipment.

Incentives: through blockchain? Biobanks. Search tool (is there, Morris!)

Metadata standardization. RING studies. Sociotechnical solutions. Money to be spent on agreement. Improvement of curation over time.

Third group: data access plus authorization mechanisms. Data quality checks, part of the curation process, should become part of the metadata. Bottom-up governance. Availability of tools.

Already there: workflow systems (Galaxy). PHT, OpenRISKNET, DTL and ELIXIR networks

To be developed: incentives for making data available, plus curation efforts; acceptable part of a thesis (instead of a chapter).

Contributions: tools that bring bioinformatics to the wetlab.

Fourth group (Alida): Easy access, easy answer to recurring questions: user-driven. Metadata on measurements and samples. Meta

Annotation tool there

Docker-based installations of tools and workflows. Plug and play.

Finances for keeping up data and services after the funding of a project. KWF and ZON-MW are already thinking about this.

What needs to be developed: backend platform with API. Standardized authorization mechanisms.

Legal side: data ownership; c

Funding budget for maintaining data and services.

Search engine for data.

Proteomics: update community

Cloud integration. Vendors: universal metadata. Inexpensive services by companies.

Conclusions:

Recurrent

We do not know what is out there!

We do not have time: incentives needed. Token engineering. Reward beyond the publications. Start counting;

How should DTL spent time / expenses

Minimal viable products (we have those). We do not know from each other!

A new search engine. Google data search engine is there?

Be prepared for deep learning.

Personal Health Train

Varian took over the distributed learning IP from Siemens and Wolfgang Wiessler. Inga is FINDable.

Varian in the radiation therapy field (linear accelerators; based in Palo Alto).

Varian learning portal. Privacy-preserving, distrib uted machine learning infrastructure. 11 centers in 5 countries. Supports EXE, JAVA , MATLAB, R, Python. E.g, distributed SVM.

VLP Gateway in the cloud. UI, project and site management, algorithm management, learning API. IP whitelists, TLS-based encryption, Restricted communication channels, Multi-factor authentication.

Varian learning connector. Deploys and invokes algorithms.

20k challenge: 20,000+ lung cancer patients (involvement of Maastro, Radboud, ErasmusMC).

SAGE dashboard: non-technical UI.

Data responsibility (in terms of GDPR): also a distributed algorithm is co-responsible as it is seen as accessing the data.

From research to diagnosis; model libraries in app store

Wouter Franke (ZIN): how does it help us? A simulation around IAT (inter-arterial thrombectomy in treatment of stroke) (ambulance times, transfer from regional hospitals to 12 hospitals providing the treatment). FAIR data points (RDF stores), Docker, RESTful interfaces, sparql endpoints. Value proposition.

Breakout sessions:

Kees van Bochove (The Hyve): FAIR already. PHT starting up.

Mattijs Brouwer (WUR): FARM data train. Fast mover.

Marius Monen (TUE): data science center program manager (representing colleague on health data analytics). How technology ready are we?

Martine Bakker (RIVM): risk assessment of substances. PHT: is it applicable to look into different databases available for substances and nanomaterials.

Gerda Meijboom (NICTIZ): interoperability between health care providers. What is in there for general health care.

Annika Jacobsen (LUMC): making rare disease patient registries FAIR. Developing the implementation.

Ronald Cornet (AMC): semantic interoperability of clinical records. Interested to implement these in some of the FAIR registries. From FAIR principles to FAIR projects. Vaporware (from steam train to highspeed train).

Eefje Poppelaars (UL – Salzburg, PhD in neuroscience, data science to-be): interested in career option.

Filip Pattyn (Ontoforce) – Belgian company. Combining external with internal data. Can our technology be aligned with the PHT technologies.

What is the value for the different organizations?

Value for patients, how to improve care, end-beneficiary. More profound research.

Access to medical data.

Underlying infrastructure behind the learning health care system.

Optimize design of clinical trials. Real value?

Efficient reuse of data for prediction of risks of new materials

Prevention and population health. Exchange data between hospitals

Benefit from data reuse.

Value proposition for society (including patients), everyone that is in the chain. Complex example of chain computarization. Align propositions. Plus funder. European level.