

## **DTL FAIR DATA Stewardship meeting for Hotel Managers**

**Utrecht, March 22<sup>nd</sup>, SURFSARA**

---

### **Conclusions and follow-up of the meeting**

- Presentations of the speakers and a summary of the meeting will be made available to all participants and on the DTL website
- DTL will publish on its website good and bad examples of data management, and will gladly accept stories from others
- DTL will make examples of good Data Management plans available on its website
- DTL will start an inventory of tools (per technology) that are currently being used for data management and publish this on its website

### **Summary of the presentations, discussion and list of participants**

#### **Gabino Sanchez Perez (Wageningen UR)**

The output of NGS data grows (very) rapidly, generating now terrabytes per run.

This poses questions to hotel managers (actually data managers!) for instance:

- Which data to keep: raw data, intermediate data or just the end result
- How long should data be stored?
- What type of storage is preferred?
- Safety & Security of data
- Data formats & compression

Another challenge is to deal with new standards, as often these are not established (yet), especially for innovative techniques.

How to smooth FAIR transition for hotel guests

- Share experiences with PIs
- Show cases of good and bad practices
- Provide templates, but with the active need of PI's input
- IP issues – timeline for public release
- PPPs – companies involved – benefits of sharing
- Why FAIRing data will be feedback loop for companies?

#### **Thomas Hankemeier (Leiden University)**

- Assuring traceability is key: making sure that the results that we deliver now to clients can be proven and explained, also five years from now
- Proper data management should facilitate research based on (existing) research data
- An (easy to use) exchange format, using controlled vocabularies/ontologies gives certainty about what was measured and how it was measured
- Researchers need to share all information required to reproduce their results (<https://biosharing.org/pages/about/>), which means sharing:
  - SOPs
  - Scripts/software to (pre-)process the data
  - How decisions were made, for instance why data were discarded

- Workflow and data management are crucial for each facility and field
- Share good practices
- For metabolomics: absolute concentrations are key
  - Benefits for validation and replication
  - Benefits could be gained by setting up a limited number of core facilities for high throughput
- Benefits are to gained in omics integration; NL/DTL could lead by example
- Sharing of metadata is often the bottleneck

### **Rob Hooft (Dutch Techcentre for Life Sciences)**

FAIR principles

*Findable*, In which repository are data stored:

- Domain specific
- Institutional
- National
- Special
- Persistent Identifier
- Catalogue
- Repository

*Accessible*

- Longevity
- Legal conditions
- Embargo
- ICT

*Interoperable*

- Format
- Terminology

*Reusable*

- Provenance
- Minimal metadata
- License
- Not narrated

### **Joost van Kempen (Radboud UMC)**

Data management issues:

- “Researchers loose significant **start-up-time** on organizing their data infrastructure”
- “Researchers cannot **access data** in an easy and compliant way”
- “**Collecting and monitoring** data is a major effort based on duplicate and manual tasks”
- “Researchers cannot **reproduce** their **analysis** in a structured and easy way”

- “Researchers almost always work in studies across institutions, yet **sharing** data and analysis algorithms is very hard.”
- “**Re-use** of data or **retrospective** studies often not even on the radar as it is virtually impossible to do.”

Radboud UMC is creating a Digital Research Environment (in the context of the NFU

Data4LifeSciences programme) to enable its researchers to work according to the FAIR principles

- All data and analytics is **findable** for everyone who is allowed to see it
- All data is **accessible** for everyone who is allowed to access it
- All data is as **interoperable** as feasible as result of the used standards
- All data is **annotated** for everyone who is allowed to use it
- Supporting the life science community using a **federal approach**
- **Integral approach** of research and care
- **Integral data management** in a single cohesive and compliant environment
- Supporting **standards** and best practices
- **Self-service** and **unburdening** (compliance on the background)
- **Flexible** and **scalable** (also small studies with small budgets)

## Discussion

- Besides proper data stewardship, data quality is equally important, making sure that data are correct (and reproducible!), so that sharing makes sense
- How do we get people in universities and biobanks interested in FAIR data and data stewardship? What is the incentive/advantage to participate (carrot) instead of just telling that it must be done
- Find examples why bad data stewardship is really hurting, and examples where combining data leads to (much) better results
- Clear formats are lacking in some instances, making data stewardship really a pain in the neck. For instance for experimental set-up.
- Companies involved in PPPs have difficulty in sharing their data in public.
- Putting data stewardship in practice is difficult (Proteomics Wageningen). Especially logging of the practical information during the experiment and easy tools to do good FAIR data stewardship
- Would a LIMS system be better implemented, that would greatly help. Having a good organized data flow in your organization helps in organizing your research.
- Coupling of e-labjournal or One Note is still not easy. ISA TAB tools (developed in Oxford).
- Sharing information on what tools to use would be extremely beneficial. They are also being developed in the framework of PHENOMENAL. That could also being taken up by the Programmers meeting of the people doing the work meet and could exchange. We need to make an INVENTORY, with PROs and CONs. Link with TRAIT and Lygature (action for DTL)
- We could organize another FOCUS meeting on this.

**List of participants:**

		Organisation
Twan	America	Wageningen UR
Ke	Lin	Rijkswaas Zaden
Thang	Pham	OncoProteomics Laboratory
Bas	Van Breukelen	Utrecht University
joost	van Kempen	Radboudumc
Xiangyu	rao	Novogene Bioinformatics Technology Co., Ltd
Marc	Broekhoven	TU Eindhoven
Alain	van Gool	DTL
Thomas	Hankemeier	Leiden Universiteit
Gabino	Sanchez Perez	Wageningen UR
Merlijn	van Rijswijk	DTL
Rob	Hoofst	DTL
Twan	America	Wageningen UR
Roland	Mumm	Plant Research Int., Wageningen UR
Jantine	Dirksen	Quaero Systems
Gerrit	Polder	Wageningen UR
Hans	Wienk	Utrecht University