Report of the DTL focus meeting on Life Science Data Repositories

Goal

The goal of the meeting was to inform and discuss research data repositories for life sciences. The big data era adds to the complexity of data environments and creates challenges in data management, stewardship, and resources to ensure data accessibility, stability and reliability. By all means research data repositories become an increasingly essential infrastructure component.

The cooperation between partners within DTL will help to develop (an ecosystem of) data repositories that suits the requirements of the life science and health research community, including data sharing and exchange, data use and re-use, data discovery, stewardship and long-term preservation.

Setting

The setting of the DTL focus meeting was informal, with presentations and room for discussion based on the presentations as well as between participants during the breaks. Participants included a mix of DTL researchers (largest group) and bioinformaticians, librarians, data repository / archiving service providers, and IT managers.

Program

The progam started with a session on data repository solutions from other sciences and EU.

- Madeleine de Smaele, 3TUs: research data archiving
- Marjan Grootveld, KNAW-DANS: data archiving and networked services
- Willem Elbers (due to illness talk presented by Irene Nooren): EUDAT: European data infrastructure
- Niek Bosch, SURFsara: research data storage and sharing

This was followed by a session on data management & repositories in the life sciences:

- Leon Mei & Morris Swertz (LUMC, BBMRI): current status & requirements genomics data
- Hugo Vrenken (VUmc, SNID workgroup): sharing neuro-imaging data
- Discussions & conclusions. Breakout sessions data life cycle themes:
    - Data design & planning (Irene Nooren)
    - Data capture, processing & curation (Morris Swertz)
    - Data analysis & integration (Jildau Bouwman)
    - Information & insight (Katy Wolstencroft)

Knowledge exchange

In the introduction the data life cycle that DTL practices was explained as well as a definition of a data repository. Fundamental to the data repository is the data

storage. The repository architecture manages content data as well as metadata, and offers a minimum set of basic services e.g. put, get, search, access control. The repository must be sustainable and trusted, well-supported and well-managed**.** Characteristics are: Sustainability, security and trust, accessibility, reproducibility, discoverability and usability. Solutions require long-term preservation, authentication and authorization, data governance, interoperability and meta data management. Currently, several initiatives are ongoing to draw up a data stewardship plan incorporating these aspects

In view of life science research, data management is complex because of its heterogeneous data. Various domain-specific data management systems and portals are currently available, either institutional, or shared within a community. It concerns different data types and complex data models, and includes up to Terabytes of data.

The first session of the program included talks from organizations within the Netherlands that provide data repositories.

- 3TU.Datacentrum. The goal of 3TU.Datacentrum s is to provide long term access to valuable research data, discoverable and usable data collections, to reskill researchers and support staff, to provide access to tools for management of research data and to offer support services for research data management. 3TU.Datacentrum currently provide these services mostly for the earth sciences, applied sciences, and disciplines in technology and construction. 3TU.Datacentrum is specialized in measurement-, geo- data and software.
- DANS. The mission of DANS is to promote and provide permanent access to digital research information. The NARCIS portal[1], managed by DANS, provides access to scientific data, people and publications held in currently forty institutional repositories. The DANS electronic archiving system[2] currently holds 257 records on life sciences amidst 26,000 datasets from humanities and social sciences. The archive carries the international Data Seal of Approval[3]. DANS and 3TU.Datacentrum collaborate as Research Data Netherlands on the Dutch Data Prize, offering training to data managers, and developing common long-term data services that are aligned with the University front offices (University Libraries and Research Support).
- EUDAT is a project that builds towards a pan-European Collaborative Data Infrastructure, driven by community requirements. 25 European partners are involved including national infrastructure providers as well as research consortia like the Virtual Physciological Human. It provides 4

[1] NARCIS: http://www.narcis.nl/

[2] DANS archive: https://easy.dans.knaw.nl/

[3]The data seal of approval are based on the following criteria:

- The data can be found on the internet
- The data are accessible (clear rights and licenses)
- The data are in a usable format
- The data are reliable
- The data are identified in a unique and persistent way so that they can be referred to

The 16 guidelines can be found at
http://www.datasealofapproval.org/en/information/guidelines/

service layers: B2SAFE: replicate research data safely, B2STAGE: get data to compute facilities, B2SHARE: store and share research data, and B2SAFE: replicate research data safely. Collaboration projects are set up to include more use cases and build more domain-specific meta data rules.
- SURFsara currently provides data storage facilities and products like short time data storage on HPC Cloud, BeeHub and the central archiving system, all within the data preservation layer. The content and context preservation layer are however important in managing metadata and subsequently understand the relevance of the data. SURFsara currently implements all layers using EUDAT technology to provide services for the complete data life cycle.

The second session summarized the current status on life science data management:
- Sharing genomics data is still a technical and organizational challenge. Leon Mei presented the LOVD database that manages locus specific mutations. Finding funding to sustain the database. A way forward could be to look for collaboration with ELIXIR, e.g. EBI for adopting and sustaining software and data management projects. Café Variome and EBI-EGA provide similar toolings. The challenge is hospital firewalls, policies and privacy issues. Even pseudoanonymization is problematics, due to lack of regulations.
- The BBMRI Biobank-based integrative omics studies (BIOS) require data integration of various data items like expression data, methylation, and genome-wide genotype data. Data is currently centrally stored on national storage facilities. Integration of data on the level of metadata is currently managed in CouchDB. A collaboration project with EUDAT has been submitted to further establish data integration and metadata management using also the Fairport concept.
- The genomics coordination centre of the University of Groningen (RUG) deals with the analysis of lots of different type of data items like NGS DNA and RNA, QTLs, and metabolomics. The design of a experiment including the data analysis is a group process that involves initiation, planning and execution. Toolings to do this have been developed in Molgenis, e.g. annotations are done using self-describing data formats, sample and analysis tracking sheets for biobanking and standard operation procedures (SOPs) for data analysis processes.
- The Sharing Neuroimaging Data (SNID) workgroup (NWO supported) with members from Dutch UMCs and Universities aims to produce guidelines for sharing of neuroimaging data including MRI, EEG and CT scans in terms of technical, financial, ethical and legal aspects. Questions that reveal are: is there a informed consent for sharing always necessary? Who is responsible for data integrity and meta-data standardization? And how to reward researchers and institutes for proper data management and sharing. Interviews have been held with stakeholders, and further discussion will take place at a Lorentz workshop, August 2014.

Discussion

A discussion was held on 4 different topics with several questions with groups of 8-10 people. The input below was given by the group of people involved in the discussions.

Data design & planning

- Who should take the lead in preparing data management plans in research? It was agreed that this is the responsibility of the funding agency.*
- Who should pay for data management work? Part of project finances should be allocated for this.
- Who is responsible for proper data management? It should be a local scientific committee from the research consortium or institute that appoints a teamleader to do this, that can or should be able to make use of legal advise.
- How do we get commitment and awareness from scientists to work with a data management plan? We should be able to give them credits. A rewarding model should be created. And things should be made as easy a possible for the data scientist, i.e. proper facilities and tools.
- How do you know where to find the right data repository? Which repository is trusted / sustainable? The funding agency should be able to provide options of trusted parties to choose from, i.e. a long list of trusted digital repositories.
- With unlimited storage facilities you would want to keep all data, but this is not realistic. What data do you store? Intermediate and/or result data? This depends on the project. In principle, results should be reproducible with the raw data and data processing or analysis workflow. There is a time issue, as for some projects it takes weeks to reconstruct the output data. In these cases you may choose to also store the result data. If referred to in a publication, result data should be stored as well.
- Where do get help on experimental planning? Ideally there is a local office with experts within the institute. It is important that these local groups share expertise, e.g. nationally, pan-Europe.
- What are the most important topics in a data management plan? When to discard data is an important question. It was also mentioned that a data management plan is a working document, not static. When research progresses, you should be able to adapt the data management plan for new data that is generated.

* During the plenary discussion it was mentioned that the institute should take this responsibility.

Data capture, processing & curation

- For repositories to be useful you need a domain-specific window into the data, user interface.
- What works well? Array express, and others were mentioned.
- What you need? Methods to define meta data, community to what they need per domain, capabilities and tools.

Data analysis & integration

On data analysis

- Do you need data sources for your data analysis (workflow)? Yes data sources are needed. However, copyright is not-known, which makes it hard to use several sources automatically. In the sources the copyright information should be made machine readable and standardized This point should be made it part of project planning (open science group of open knowledge foundation
- Are these sources machine readable? DCC site, but too many and changes,
- Do you have enough capacity for you analysis (e.g. storage)? Buy enough, cloud-solutions (security is an issue), consider laws (mega.co.nz is encripted) Separation long (sharing) and short-term (analyse) data repository awareness.
- Do you keep track of your analysis versions? Which steps taken is also data à audit trail needed (needed to automate pipeline), include versioning of scripts used, automatic lab notebooks (use Galaxy, Katy Wolstencroft, myExperiment.org, Paul Groth)
- Are those interoperable? Use meta-data standards to make exchange possible (make machine readable meta-metadata)

On data integration

- Is the data you want to integrate in the same format?
- Do you need to process your data with the same workflow?
- Is the data you want to integrate in the same unit?
- Can you connect the identifiers related to your data (semantic interoperability)?
- Is there a statistical solution to integrate your data?
- Do you need data sources for your data integration?
- Are these sources machine readable?
- Are these sources public?
- Do you have examples for (data) repositories for this purpose?

The answer to these questions it to make data comparable by using semantics and meta-metadata (RDF). Make data exchange possible by using API's (e.g. SOAP).

Other questions that came up during the meeting:

- Value of data sets relative to each other: Store all data, user pays at least for 5 years, after that user of data pays, duplication will be important here
- Negative results? Lost? Should be part of planning, separate from useless data
- Citing subsets of data? Use nanopublications

Information & insight

The discussion was driven by use cases in medical imaging.
Specific issues

- Data are not always shared, even after publication. At most, only data featured in publications is shared.
- More and more difficult to make data anonymous

- Once informed consent has been given, and data has been shared, it is very difficult for patients to change their minds. This is a barrier to the initial consent.

What we need

- A local catalogue of data would be useful for keeping research records and allowing reuse with immediate collaborators.
- Different requirements for data publication.
- Ensure people make their data available to authorised people (i.e. other researchers).
- Ensure patients can trust the processes of consent and anonymisation.

Incentives

- Funding agencies can enforce data sharing policies and develop new policies where they don't exist (medical imaging practices are not the same as in other life science areas)
- Publication credit. Datasets should be citable. Some consortia offer data for reuse in return for co-authorship
- For patient data, put them in control. They decide when their data is shared and reused. This is a scheme that is already being implemented in the USA.

Conclusions

The DTL focus meeting lead to lively discussions due a good mix of participants from different areas and backgrounds, e.g. librarians and researchers. Clearly, data management and repositories is a field that requires development with involvement of many parties. Funding and commitment of institutes to provide data scientist roles is fundamental to this development. The life science research landscape is heterogeneous with a widespread ecosystem of data management systems. Researchers that are in charge of data need to be reached and facilitated with tools and repositories to manage their data. Further development and focus should be on the technology domains of annotation and interoperability, data provenance and metadata management.