Bring Your Own Data Workshop - OncoXL

Evaluation report

Author: Daphne van Beek
Date: June 27th, 2017

Summary

This workshop was a good proof-of-concept of the application of the FAIR principles and the resulting deliverables can be used as a good example of what can be achieved if you publish your data in a FAIR manner. To use the dataset as an example; deployment of the FDP is still one step that needs to be made. To further increase awareness, publishing a paper on the workshops progress is still under discussion.

Background and scope

Cancer **G**enomi**C**s.**nl** is a consortium of research groups from the University Medical Centre Utrecht, the Hubrecht Institute in Utrecht, Utrecht University, the Netherlands Cancer Institute in Amsterdam, the Erasmus Medical Centre in Rotterdam, the Academic Medical Center Amsterdam, Leiden University Medical Centre and Radboud University Medical Center Nijmegen.
Within a CGC research project, samples were analysed using different omics techniques, resulting in the draft paper 'A multi-omics approach identifies widespread responses to synergistic BRAF and EGFR inhibition in CRC cells'. During this project, the omics results were combined using in-house scripts. Resulting datasets were not interoperable at this stage.
During the 'Bring Your Own Data Workshop - OncoXL', the goal was to completely FAIRify the four different omics datasets available.

To obtain a completely FAIR dataset, input is needed from both linked-data experts, as well as domain experts. The workshop resulted in four spin-off groups, surrounding one or multiple domain experts per dataset:

- DNA-seq
- RNA-seq
- Proteomics and Phosphoproteomics
- Experiment

After three days of hands-on sessions, evaluation on the workshop took place.

Deliverables

The main deliverable at the end of the three day hands-on workshop, are four FAIR datasets, one for each omics domain, which are interoperable with each other, as well as with external sources.

This main deliverable could be defined in smaller deliverables:

1. **Data models for each omics-field**: a data model is a representation of the data that is contained within the dataset. A good model can be (partly) reused and linked to other data models if a node overlaps. The model can be designed in such a way that integration with large, public datasets will become a possibility.
2. **(FAIRifier) script for transferring the raw dataset into the RDF representation**: after the development of a data model, the dataset itself can be transferred into a FAIR dataset by applying the model on the data. This requires a script or FAIRifier protocol. The result is a RDF Turtle file that can be stored in a triple-store and queried.
3. **Knowledge about the ontologies that are available within the different fields**: to generate a good working data model, knowledge about different ontologies and experience in making data models is important. This workshop is the first step in creation of this knowledge and experience for the participants.
4. **Metadata at different levels for generation of the FAIR data point**: a FAIR data point is an 'access point' where users can take a look at the data. To make sure the data can be found easily and specify the requirements for reuse, a good set of metadata should be developed. In this case, there are multiple metadata levels: catalog, dataset and distribution levels.
5. **Deployment of a FAIR data point**: The final step is the publication of the dataset as an access point: the FAIR data point. This data point contains the raw data files and the RDF representations of the files as distributions and the dataset metadata at different levels. It can be indexed by search engines, making the data findable.
6. **Slide-deck**: containing our recommendations and conclusions.

Evaluation

The workshop started with an introduction into FAIR principles and available tools and the division of the participants into working groups. The groups each went their own way in producing the deliverables. For each group, there was at least one linked-data expert available that could assist in the transformation of each data set.

On the first day, the main research question was defined: "*At time point X what are the 'all types of' values for each condition and omics level for protein Y*".

Each group produced deliverable 1, the data model. While every participant went their own way, quickly it became clear that parts of the model were indeed overlapping between datasets. It only took a little tweaking to get the models overlapping by choosing the same ontologies. The group that was dealing with the proteomics and phosphoproteomics eventually focussed on the proteomics dataset, as the datasets were not as similar as was previously expected. The data models are made available at http://www.ubec.nl/data/fair-data-point/.

Halfway past the second day, the groups started to apply the models on the datasets, all using different methods:

1. Using the FAIRifier tool to apply the model (proteomics)
2. Writing a custom python script (DNA-seq, code can be found at https://github.com/UMCUGenetics/FairDNA)
3. Writing a custom python script that loads the data into a template (RNA-seq, code can be found at https://github.com/UBEC/RDFtemplates). This setup can be reused and could be modified to work with any input RNA-seq data.

All methods produced a valid RDF Turtle file that could be queried using SPARQL.
The experiment group created a model, but did not transfer the model into code. They did generate a policy that resulted in unique identifiers for each sample at each time point (each point of sampling that was done in the experimental setup). These sampling-point names were then used in the Turtle files that were created by the other groups.

One of the domain specialists put time and effort in the creation of the metadata files on the levels of catalog, dataset and distribution (see figure 1 for more information about this structure).

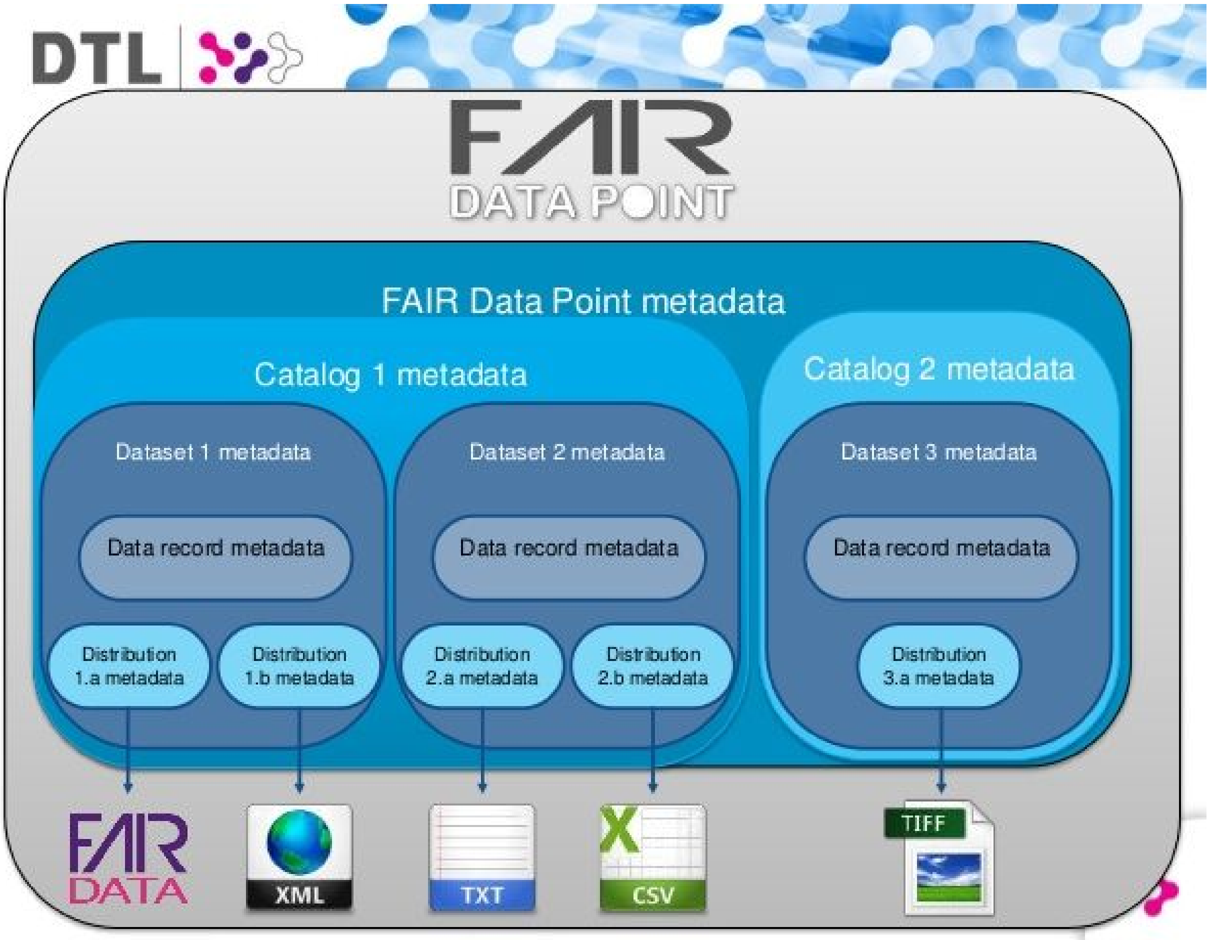

Figure 1: Representation of a FAIR datapoint, containing the different levels of metadata that should be made available. [Source]

In Figure 1 the way the metadata is structured is made clear. The description of the layers of metadata can be found in Appendix A.
The different levels of metadata for the OncoXL project were defined and produced, so they could be used in the deployment of the FDP.

For the deployment of the FDP, a server with specific requirements is needed. During the BYOD workshop, such a server was configured, but there was not enough time to actually deploy the datasets on this server. There will be an effort to still deploy the FDP, first locally at the UMCU to see if the configured server is fully functional and test the way to implement a local FDP. If this proof of concept is successful, the FDP will be transferred to a FDP that is hosted at SURF, as the project is a consortium effort and should benefit all contributing parties.

Until the deployment of a FDP, the materials produced in the workshop will be available at http://www.ubec.nl/data/fair-data-point/.

Conclusions and Recommendations

In the three day hands-on BYOD workshop, three separate datasets were completely FAIRified, using different but partly overlapping models that facilitate interoperability between the datasets.

The communication between the different workgroups went smoothly, as was needed to make the data models interlinkable (same ontologies should be used for the same concepts in the data). Composition of the groups was ideal; the ratio of linked-data experts to domain experts/bioinformaticians was good.

The definition of the research question in the beginning of the workshop was very relevant; it kept the focus on what part of the datasets were most important (instead of modelling everything and generating a too complex system).

The selection of ontologies is one of the hardest steps when creating a data model. Where should you look for relevant ontologies, which one should you choose? If there is none available, should you extend an existing ontology or create a new one yourself? These questions were raised and only partly answered due to the complexity of the problem.

- Experience in building data models will create better understanding about this problem
- Adaptation of ontologies when describing datasets will increase the production of relevant ontologies in all fields.
- A general introduction into ontologies (explaining concepts: classes and properties) is needed when starting with the application of FAIR.

One of the problems that we ran into was how to deal with ‘versioning’ of annotations and redundancy in your data. For example, in the DNA-seq VCF, there were annotations made for each variant that contained a gene name. As gene names can be changed over time, this should be taken into account (maybe not use the gene name but a unique ID, or use SPARQL queries to annotate the variant directly). How should one deal with multiple annotations (multiple genes)?

Another limiting factor was the time. There were a lot of deliverables defined, and not all of them could be created. Not all participants could keep up with the steps that were done in other teams, there was not enough time to share all the developments. A solution to this problem could be to start generation of the data model before the workshop begins, so only part of the first day has to be spent on correcting and finalizing the model with the help of the linked-data experts.

This workshop was a good proof-of-concept of the application of the FAIR principles and the resulting deliverables can be used as a good example of what can be achieved if you publish your data in a FAIR manner. To use the dataset as an example; deployment of the FDP is still one step that needs to be made. To further increase awareness, publishing a paper on the workshops progress is still under discussion.

References

Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*3:160018 doi: 10.1038/sdata.2016.18 (2016).

Mons, Barend *et al.* Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use*, vol. 37, no. 1, pp. 49-56 doi: 10.3233/ISU-170824 (2017).

Appendix A: FAIR metadata layers

See Figure 1 for an overview of the structure. The FAIR data point (FDP) should have metadata about its contents and its position. One could for example imagine that you eventually would have a FDP per institute or PI group. In the metadata you describe what the focus of the data point is (what type of research/data), the affiliation, where a person can find more information and for example information about the licensing and the way to contact a data access board. In the FDP you could have multiple (research) project that can be divided in catalogs. These catalogs contain similar information about the affiliation, licensing and data access board, but could also contain contact information of the people involved in the project, publications, ontology terms describing the specific research field, etc.

Within the project, there is space for different datasets, for example DNA-seq and RNA-seq (or a protocol in Word format and an Excel sheet that contains the raw and analysed data). The dataset can again be described in a way it could be reused by others/externals. It states the contents of the file and how it is produced (which protocol/programs), and can be extended by as much metadata as the researcher is willing to provide. The last level of metadata is on the distribution level: here you can find metadata on the different distributions of a dataset, which are effectively different formats of the data. For example, the Word document can be described here, but to improve interoperability, the contents could also be converted into a TXT file. This metadata could contain a direct link to the data, but could also point to a location that can only be accessed when authorized, or contain a protocol for a user to follow to gain access to this specific file.