Solanacea-centric Linked Data Platform for Plant Breeding & Genomics

Scientific Symposium: FAIR Data Science for Green Life Sciences

netherlands eScience center

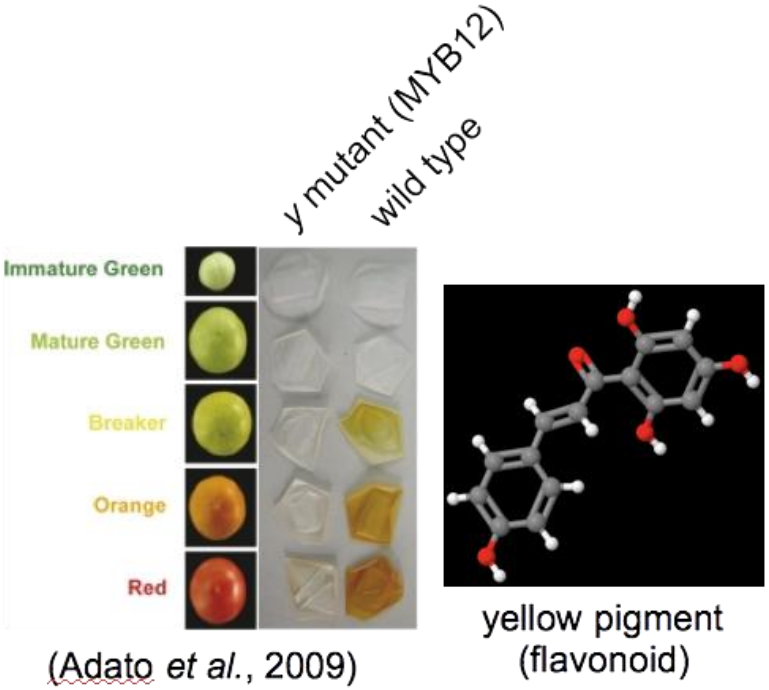
by SURF & NWO

WAGENINGEN UR
For quality of life

Arnold Kuzniar

12/12/2018

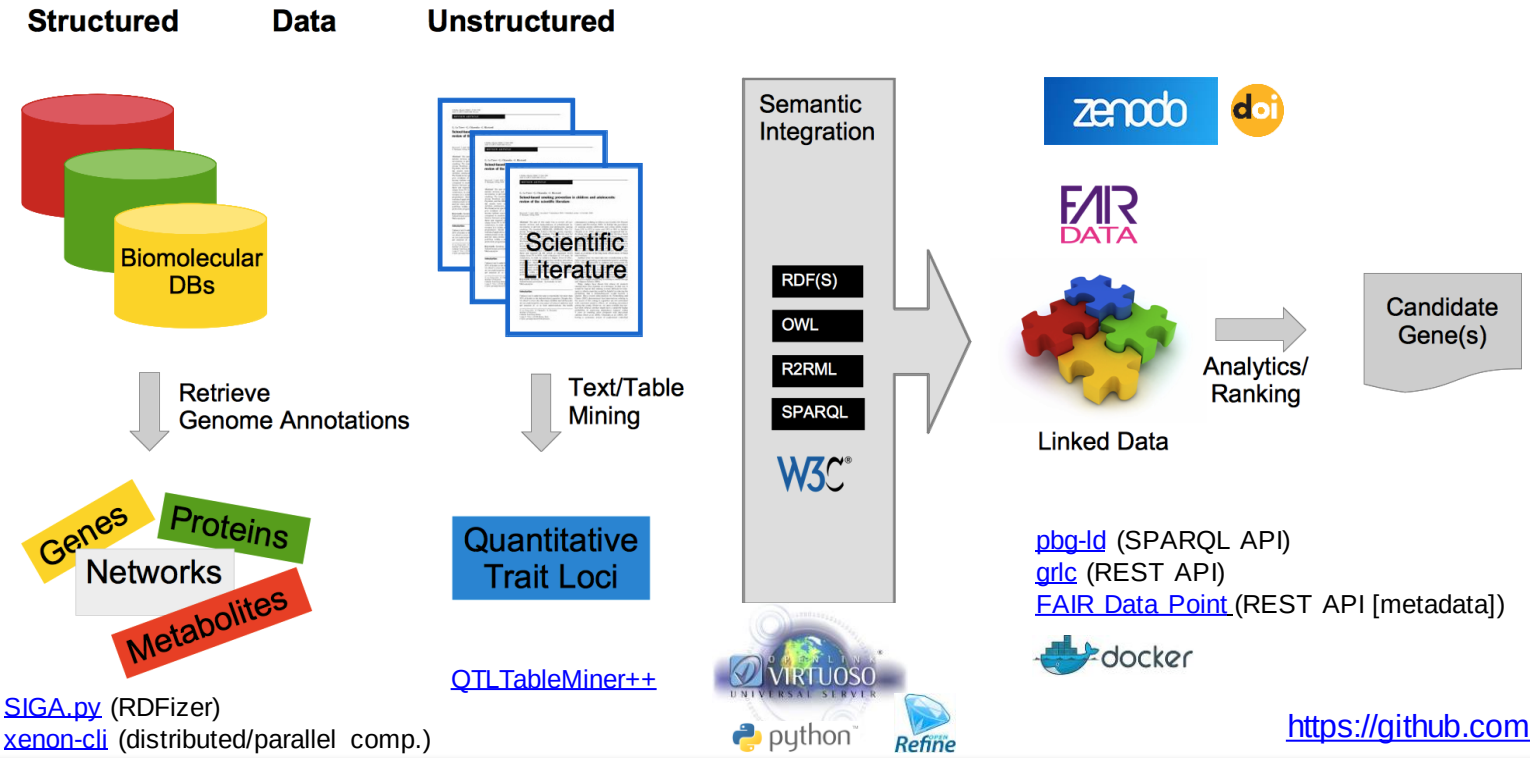
Linking genes to traits



(Adato *et al.*, 2009)

yellow pigment
(flavonoid)

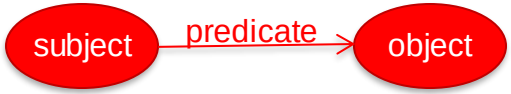
Methodology



Linked Data approach

Sir Tim Berners-Lee (W3C)

- Use (persistent) URIs/IRIs to identify things
- Use HTTP-resolvable URIs
- Use RDF graph model (triples)
- Link to other resources



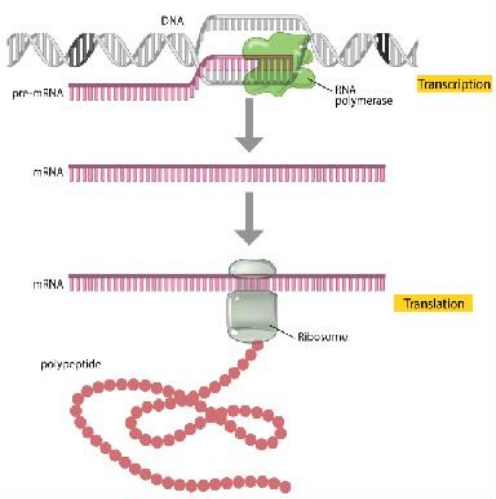
<URI> <URI> <URI | Literal>

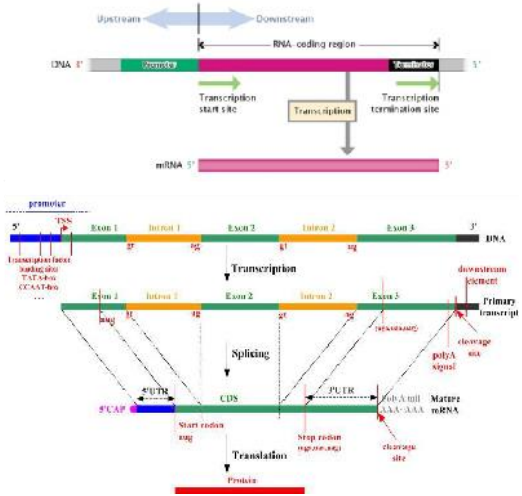
RDF turtle format

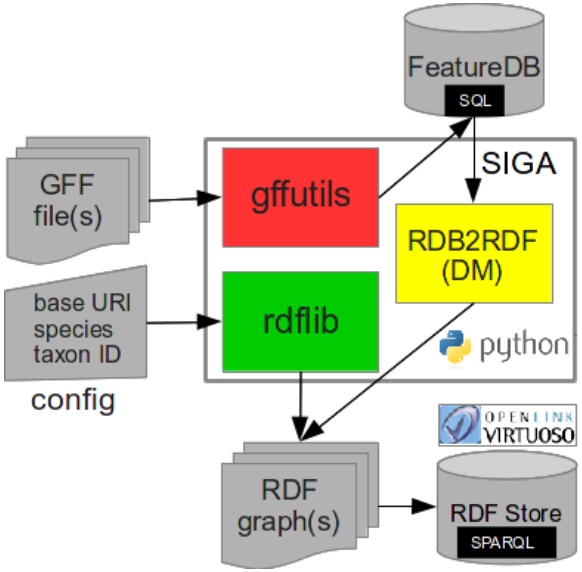
```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix obo: <http://purl.obolibrary.org/obo/> .

<http://solgenomics.net/genome/Solanum_lycopersicum/gene/Solyc00g005000.2> rdf:type obo:SO_0000704 ;
    rdfs:label "gene Solyc00g005000.2"^^xsd:string ;
```

Solyc00g005000.2 is a gene

Genome annotations



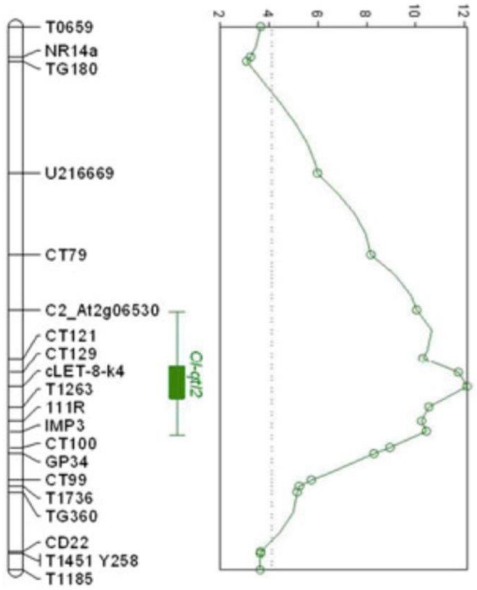




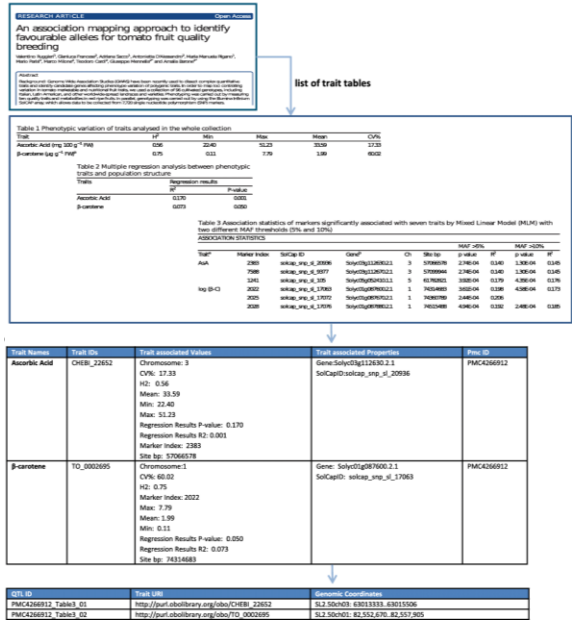
Use controlled vocabularies & ontologies
(e.g. GenBank Feature Table, SO[FA], FALDO, DCMI)

Software architecture

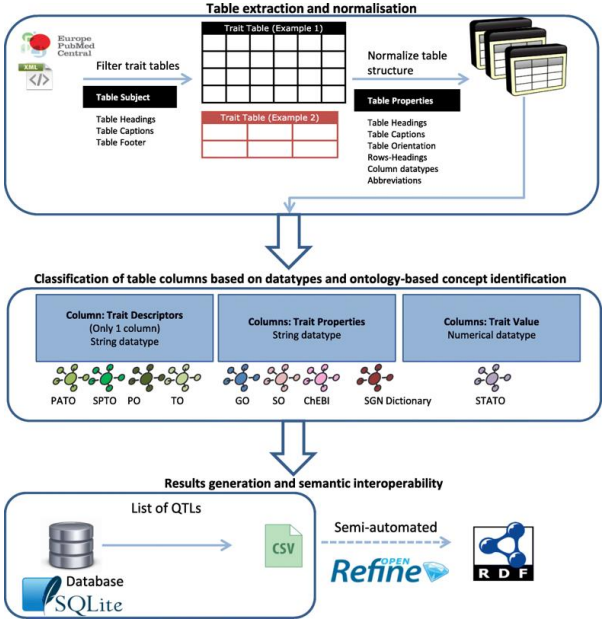
Mining QTLs from literature



e.g., QTL mapped on chr12
(Faino *et al.*, 2012)



QTLs "buried" in tables



QTL TableMiner++
(Singh *et al.* 2018)

Data sources

Domestic tomato (*Solanum lycopersicum*)

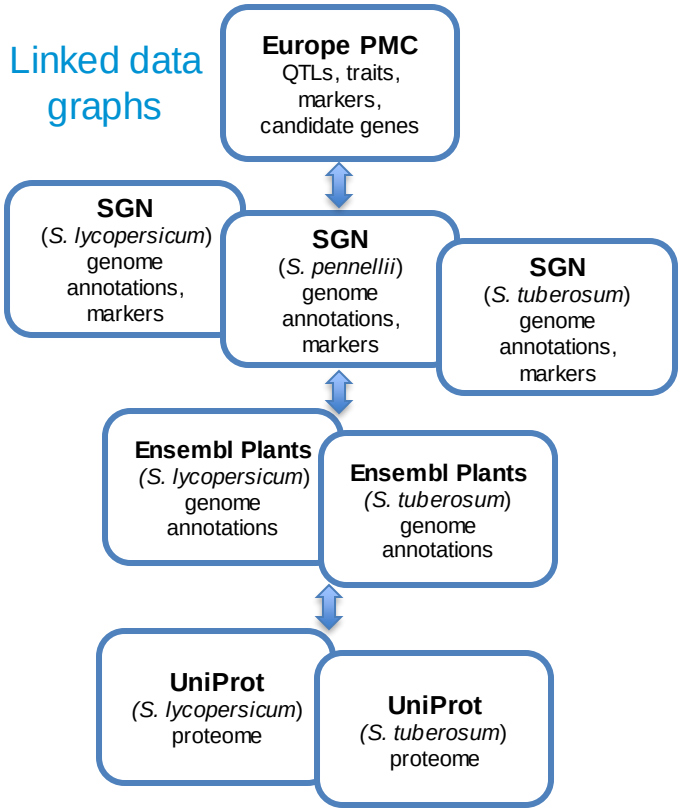
- gene models (SGN & Ensembl) + proteins (UniProt)
- genetic markers (SGN & SolCAP)
- QTLs (Europe PMC)

Wild tomato (*S. pennellii*)

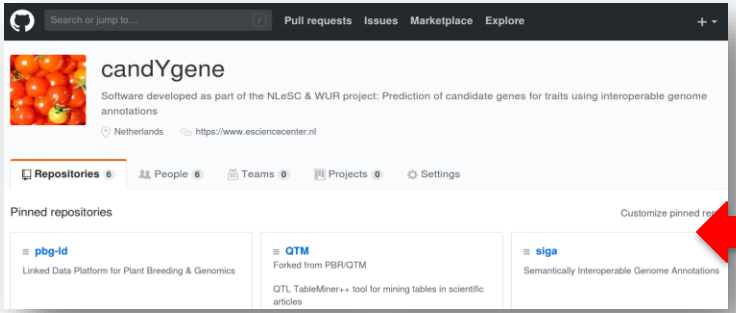
- gene models (SGN)
- genetic markers (SGN)
- QTLs (Europe PMC)

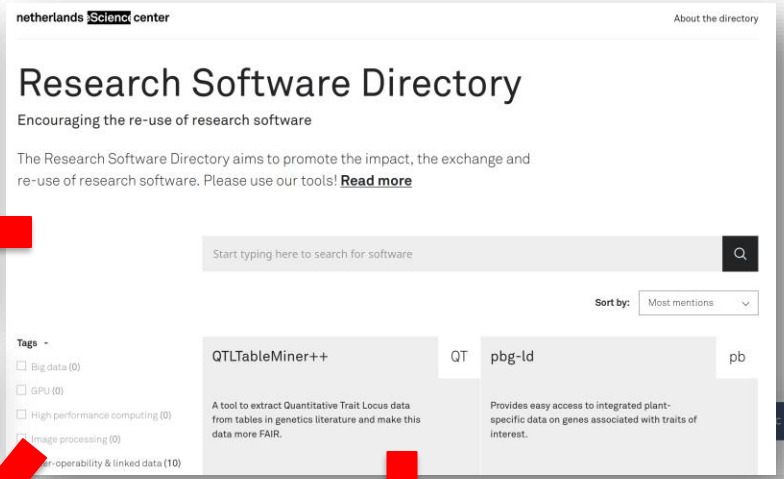
Potato (*S. tuberosum phureja*)

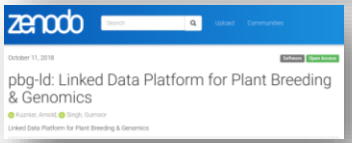
- gene models (PGSC & Ensembl) + proteins (UniProt)
- genetic markers (SolCAP & DArT)
- QTLs (Europe PMC)



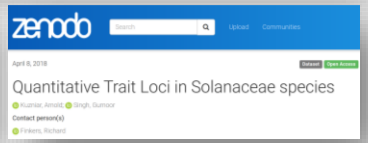
FAIR research output







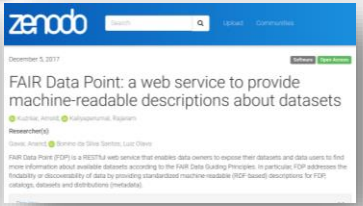
doi:10.5281/zenodo.1458169



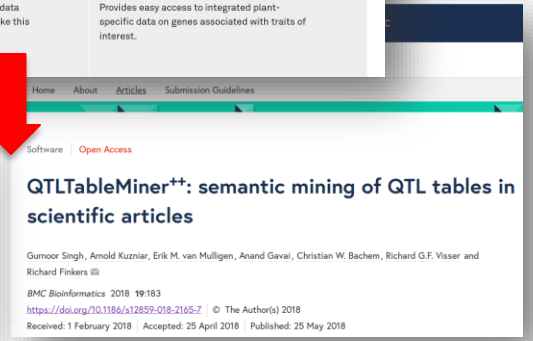
doi:10.5281/zenodo.1215044

doi:10.5281/zenodo.1193640

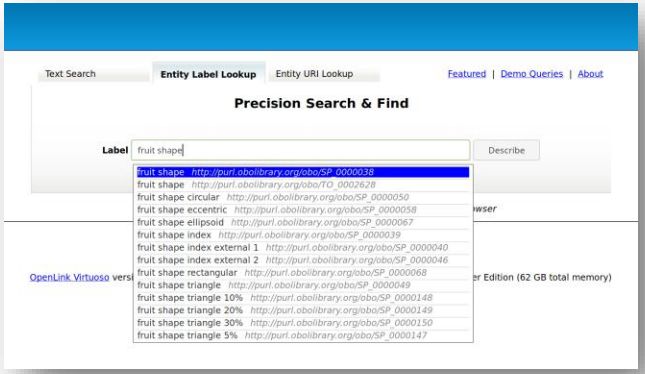
doi:10.5281/zenodo.1076438

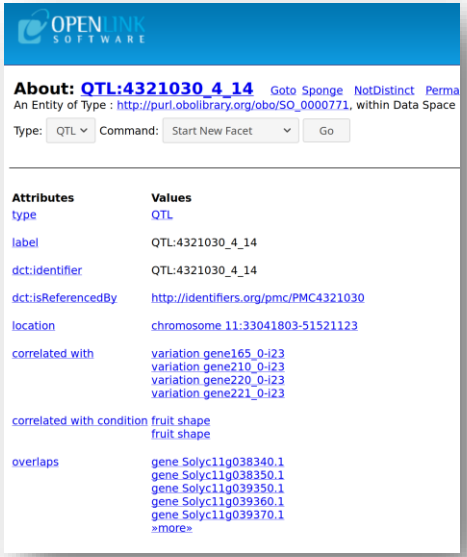


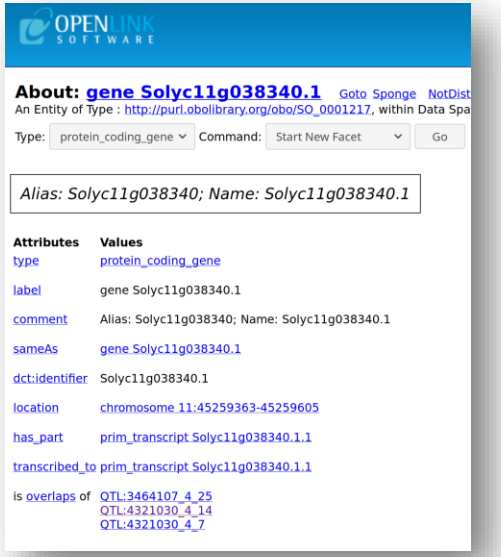
doi:10.5281/zenodo.1083951



Linked Data Platform



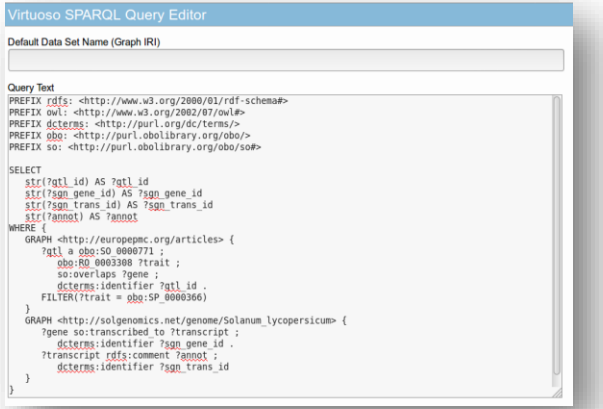




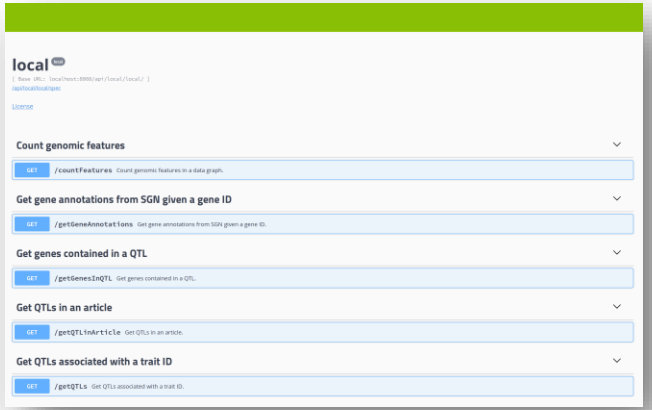
Trait ⟷ QTL ⟷ Gene

https://github.com/candYgene/pbg-ld

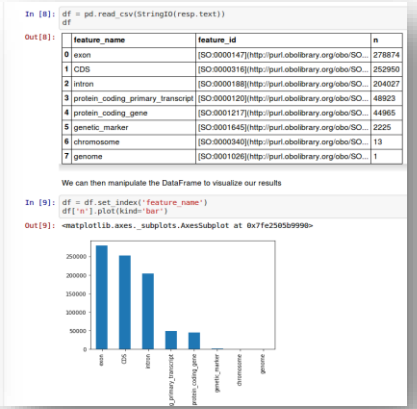
Data access & analysis



SPARQL API



REST API



Jupyter notebook

Conclusions & future work

- Use/develop open-source, standards-compliant, well-documented software
- Linked (Open) Data principles ≈ FAIR Data principles (in practice)
- FAIR compliance not only for data but also for software/workflows
- Integrating & "semantifying" (semi-)structured data requires significant time investment & domain expertise
- Improve data access & processing workflow
- Extend *pbg-ld* with algorithms for candidate gene prioritization
- Prepare new software & data releases
- Deploy *pbg-ld* software stack in production environment (HPC Cloud)

Acknowledgements

- Gurnoor Singh (WUR)
- Richard Finkers (WUR)
- Christian Bachem (WUR)
- Richard Visser (WUR)
- Erik van Mulligen (EMC)
- Anand Gavai
- Carlos Martinez-Ortiz (NLeSC)
- Lars Ridder (NLeSC)
- e-infra team (SURFsara)