netherlands eScience center

candYgene: enabling precision breeding through FAIR Data

Arnold Kuzniar[1], Anand Gavai[1], Lars Ridder[1], Luiz Olavo Bonino da Silva Santos[2], Gurnoor Singh[3], Richard Visser[3], Richard Finkers[3]

[1]Netherlands eScience Center, Science Park 140, 1098 XG Amsterdam, [2]Dutch Techcentre For Life Sciences, Catharijnesingel 54, 3511 GC Utrecht, [3]Plant Breeding, Wageningen University and Research Centre, Droevendaalsesteeg 1, 6708 PB, Wageningen, The Netherlands

Summary

Food demand is projected to increase by 50% in 2030. One approach to improve food security is by breeding new crop varieties, for example, with improved tolerance to drought or resistance to pathogens. Genetics research is focusing more and more on mining fully sequenced genomes and their annotations to identify the causal genes associated with specific traits (phenotypes) of interest. From traditional quantitative trait loci (QTLs) studies, breeders have gained insight into which genomic regions to introgress into their elite germplasms in order to improve agronomically important traits (Figure 1). However, a complex trait is typically associated with multiple QTL regions, each with hundreds of genes positively/negatively affecting the desired trait(s). Moreover, for humans it is increasingly difficult to effectively distil the growing ~omics data sets and scientific literature into relevant information and knowledge. Our aim is to develop a Big data analytics & semantic interoperability infrastructure (Figures 2 and 3) for candidate gene prioritization that will aid breeders in the design of an optimal genotype with a desired trait(s) for a given environment. Our overall goal within the NLeSC is to deliver solutions that use Semantic Web technologies to (nano)publish life sciences data in an interoperable and machine-readable form in order to enable automated reasoning over these Linked Data.

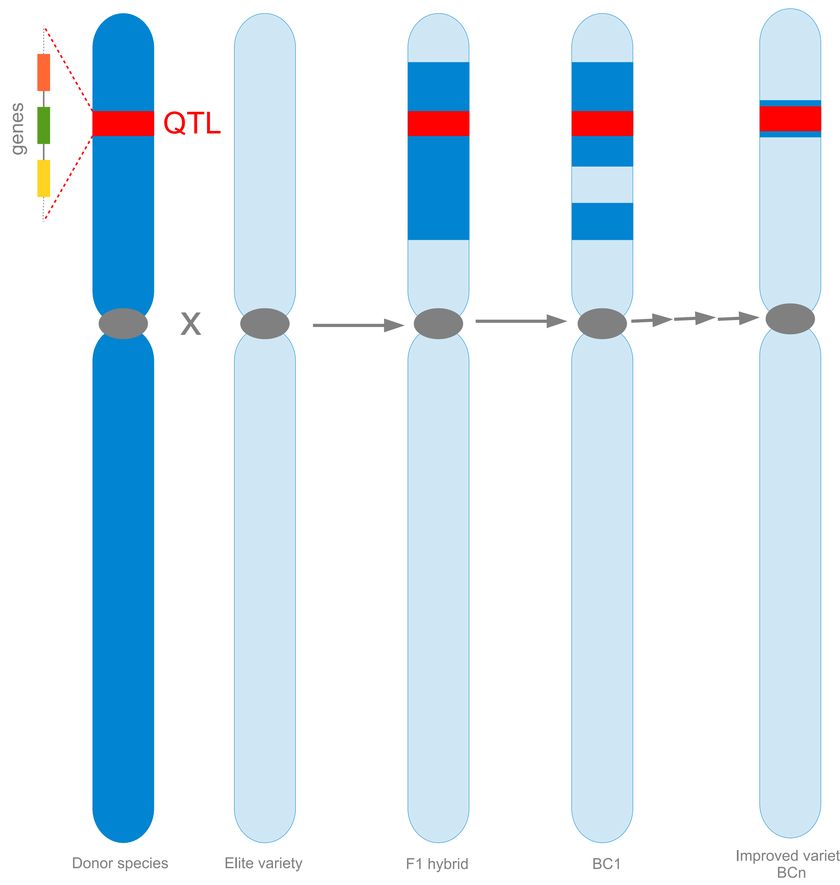

Figure 1. Schematic diagram of a common breeding scheme resulting in an improved crop variety. Each vertical bar (blue) represents the genome of an individual within a breeding population, with colored segments indicating genes/QTLs (red) associated with a trait under selection. 'X' indicates a cross between parents, resulting in the initial F1, and generation of subsequent backcross (BC1..n) populations are indicated by the arrows.

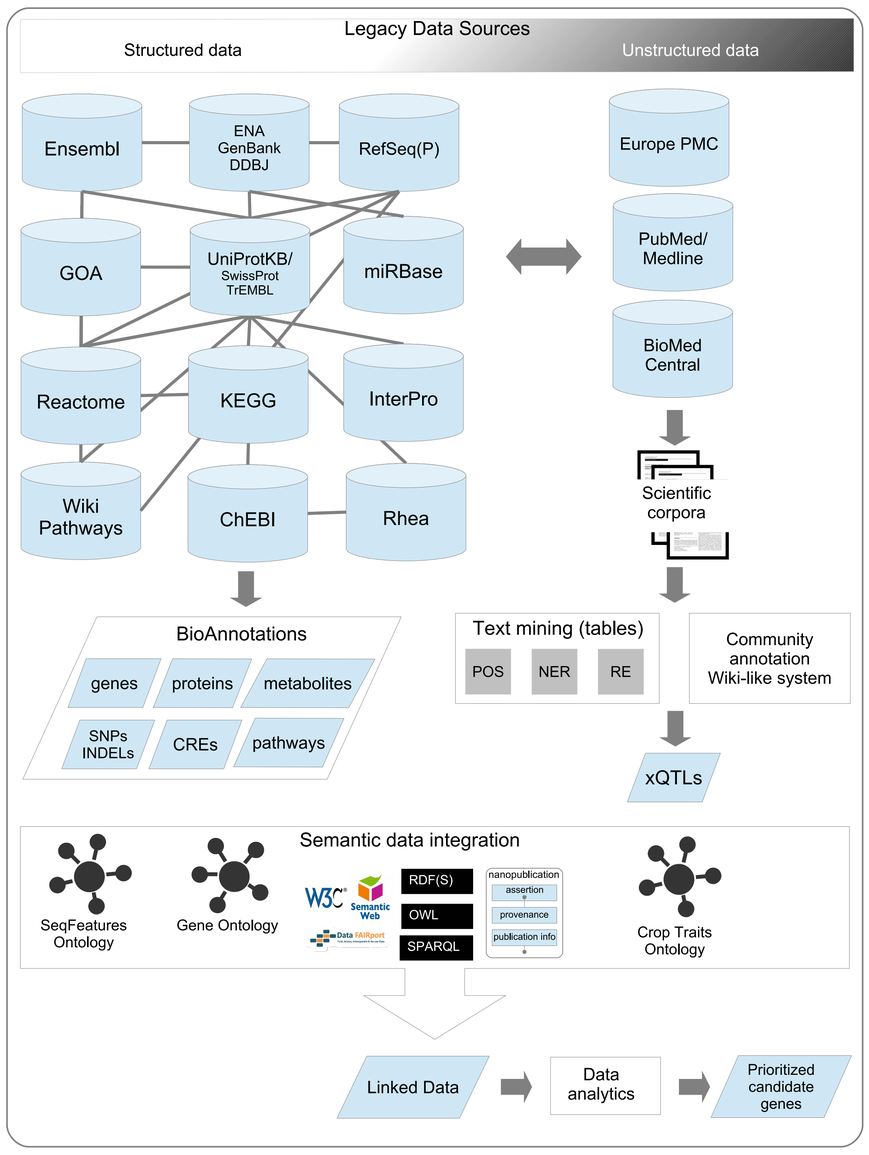

Figure 2. *candYgene* prototype architecture for mining QTLs for candidate genes. Semantic interoperability across heterogenous biological data sources is achieved by converting the genome annotations and QTLs extracted from literature into named RDF graphs with provenance and context (nanopublications [1]). A model-based prioritization will result in reduced QTL intervals with candidate genes explaining the trait of interest.

FAIR Data guiding principles

Findable	(meta)data should be uniquely and persistently identifiable
Accessible	identifiers should provide a mechanism for (meta) data access including authentication access protocol and license
Interoperable	(meta)data should be machine readable and annotated with resolvable vocabularies/ontologies
Reusable	(meta)data is sufficiently well-described to allow (semi)automated integration with other compatible data sources

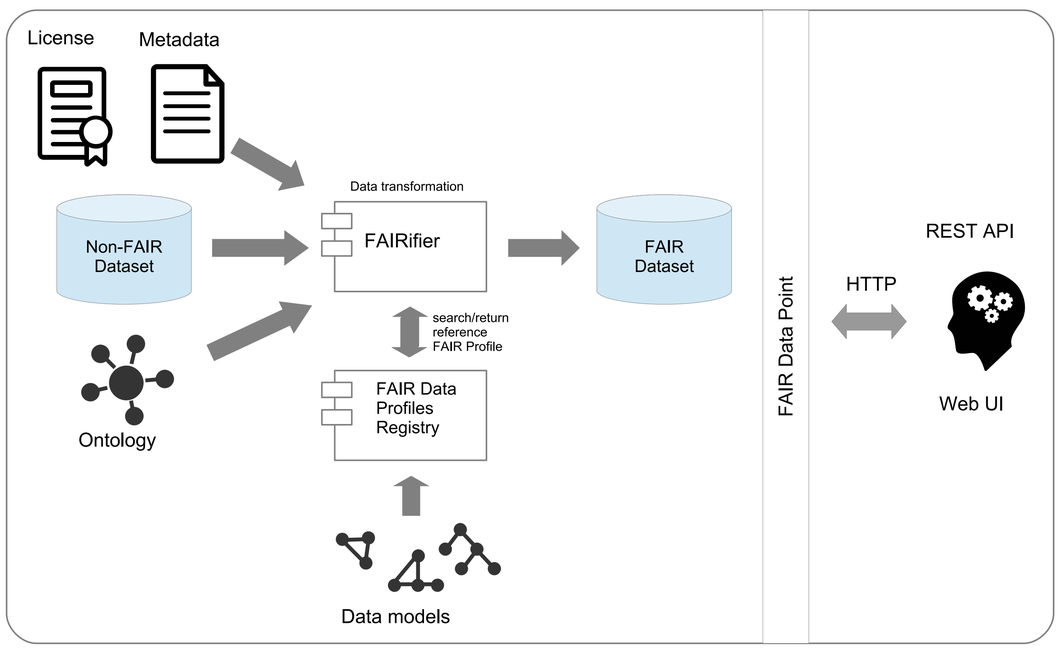

Figure 3. Prototype architecture of the FAIR Data interoperability infrastructure [2] being developed in the ODEX4all project. The prototype includes two core components: the FAIRifier consumes (meta) data from various sources and transforms these according to the FAIR Data principles (*http://www.dtls.nl/fair-data/*); and the FAIR Data Profiles Registry stores the data models underlying the datasets. A FAIR Data Point makes the FAIR dataset(s) publicly available (with access control) through a Web UI or programmatic RESTful API. One of our use cases is to make tomato passport data, provided by international gene banks [3] and private donors, FAIR compliant. This dataset includes information on more than 7000 domesticated tomato (*Solanum lycopersicum*) lines along with closely related wild species, and stored in the EU-SOL BreeDB (*https://www.eu-sol.wur.nl/*).

References

[1] Mons & Velterop (2009) Nano-publication in the e-science era. *Workshop on Semantic Web Applications In Scientific Discourse (SWASD)*.

[2] Bonino da Silva Santos *et al.* A Technological Approach for Supporting Findable, Accessible, Interoperable and Reusable Research Data. L.O. Bonino da Silva Santos *(in preparation)*.

[3] Finkers *et al.* (2015) Genebanks and genomics: how to interconnect data from both communities? *Plant Genetic Resources*, 13, 90–93.

In collaboration with ODEX4all project partners:

WAGENINGEN UNIVERSITY WAGENINGEN UR

vlpb Virtual Laboratory Plant Breeding

LUMC DTL Maastricht University Erasmus MC VU University Amsterdam FOM Radboud University Nijmegen PHORTOS

www.eScienceCenter.nl twitter.com/eScienceCenter vimeo.com/eScienceCenter