Toward better data sharing methods for genebanks

Evangelia Papoutsoglou[1], Rajaram Kaliyaperumal[2], Theo van Hintum[3], Richard G.F. Visser[1], Ioannis N. Athanasiadis[4], Richard Finkers[1]

[1] Plant Breeding, Wageningen University & Research, Wageningen 6708 PB, The Netherlands
[2] Leiden University Medical Centre, Leiden 2333 ZA, The Netherlands
[3] Centre for Genetic Resources, The Netherlands, Wageningen University & Research, Wageningen 6708 PB, The Netherlands
[4] Information Technology Group, Wageningen University & Research, Wageningen 6706 KN, The Netherlands

Abstract. The conservation of plant genetic resources (PGR) is an important task that requires collaborative effort from many stakeholders. For this, common means of data exchange and effective methods to profit from the collected information need to be established. In this paper, we describe a demonstrator promoting findability of PGR, according to the FAIR (Findable, Accessible, Interoperable & Reusable) data principles. PGR providers can each expose their germplasm information, using the FAO Multicrop Passport Descriptor (MCPD), which subsequently can be queried in a distributed manner via a single user interface. PGR users can select among predefined questions, for example for specific crops, accessions or phenotypes.. On the back end, data integration from a distributed query is achieved through annotations with the MCPD semantics.

Keywords: MCPD, FAIR, germplasm, plant genetic resources, genebanks, interoperability, data modelling, metadata, linked data, semantic web

1 Introduction

Genetic diversity in crops, and the maintenance thereof, is a crucial factor for modern breeding research. However, access to information describing this genetic diversity is not always readily available. For example, many accessions can be obtained from genebanks worldwide. Each genebank has different means to document their accessions and how to make data available, which emphasizes the need to deal with this heterogeneity. The current solution includes documenting PGR data in aggregated systems, such as EURISCO and GENESYS, however, this is not a long-term sustainable solution as the volume of information is readily increasing, especially for (~omics-derived) characterization data. We believe that a way to gather and assemble data (smaller or bigger in size and/or complexity [1]) from distributed resources will be useful, and could significantly speed up the production of results in important genomic selection, genome-wide association studies and more [2]. So, in this paper, these challenges are addressed with a demonstrator interface, relying on the reuse of existing building blocks.

2 Background

To effectively work towards a better data sharing, two aspects need to be in place. The first is a data standard to effectively describe the data. For plant genetic resources (PGR), this is the multi crop passport descriptor (MCPD) vocabulary [3]. Secondly, we need a definition on what is required to promote optimal data management/stewardship, for as example defined in the FAIR data principles [4].

2.1 Findability of PGR passport data using the FAIR data principles

The FAIR data principles dictate that all data should be Findable, Accessible, Interoperable and Reusable. Findability is crucial for the discovery of information about PGR world-wide. and requires a well-defined data standard. For example, a PGR user might want to find accessions from a specific geographical region for a specified taxa. Within the PGR community, the MCPD vocabulary is the accepted community standard to describe PGR. The MCPD comprises a set of attributes describing an accession, such as accession identifier, taxon, geographical origin, holding institute, and biological status, uniformly describing PGR. In our work, we defined a FAIR data point definition (FDP), exposing PGR data with attached metadata in a semantic manner. We will show that exposing PGR passport data according to the MCPD standard utilizing the FAIR data principles will improve findability and subsequent querying of these resources, via a query interface targeted at PGR users. The application of the FAIR principles in a demonstrator is not novel. We reuse code from the FAIR rare diseases demonstrator [5] targeted at biobanking collections, where similar questions are raised (e.g. which biobank has samples from a patient having a certain disease phenotype). This approach also shows the added value of working with diverse communities on tackling common data challenges.

2.2 Use case-relevant plant semantics resources

1. Germplasm Ontology[1]: Contains parameters not included in the MCPD, e.g. distinguishing accessions or genotype, and describing these in more detail.
2. Agronomy Ontology[2]: Defining an experiment, as a container to bind together material with other parameters (e.g. treatments, environment, etc.)
3. MIAPPE[3] (and its implementation, the Breeding Application Programming Interface - BrAPI): further describes the organization of an experiment, and holds more attributes describing it.
4. Plant Ontology[4]: generic ontology for plant structure and anatomy.
5. Plant Environment Ontology[5]: for treatments and growing conditions in plant biology experiments

1 http://www.cropontology.org/ontology/CO_010/Germplasm

2 http://www.obofoundry.org/ontology/agro.html

3 http://www.miappe.org/

4 http://purl.bioontology.org/ontology/PO

5 http://www.obofoundry.org/ontology/eo.html

6. Plant Stress Ontology[6]: for diseases and pathogens
7. Plant Trait Ontology[7]: generic ontology for the description of phenotypic traits in plants, with mappings to crop-specific trait ontologies.
8. Other species-specific ontologies, where the Trait Ontology may prove insufficient. For example, in the case of tomato, an ontology like the Solanaceae Phenotype Ontology[8] may be used for more crop-specific attributes and more agile development from that specific community.

3 Methodology and Results

3.1 Model building and choices

The main challenge was to design a model incorporating the MCPD, and attached characterization data from a (field) experiment. To do this, we identified the ontologies listed above.

Table 1. Model in terms of triples. Terms starting with a colon (:) are instances of a class, quotes (“”) enclose literal values, and brackets (<>) are used to refer to classes. Italics indicate placeholder terms, modeled specifically for this application.

#	Subject	Predicate	Object
1	:experiment_X	rdf:type	<AGRO:agricultural_experiment>
2		geo:long	“longitude”
3		geo:lat	“latitude"
4		dct:identifier	“ID”
5		dct:created	“creation date”
6		RO:has_participant	:plant_X
7		SIO:is_source_of	:observation_
8	:observation_X	rdf:type	<om:observation>
9		SIO:is_about	:plant_X
10		to:has_phenotype_score	“value”
11		to:has_phenotype_variable	:trait_X
12	:trait_X	dct:title	“trait_title”
13	:plant_X	rdf:type	<*plant*>
14		dct:identifier	“plant_ID”
15		*descendant_of*	:plant_Y
16	:plant_Y	*has_biological_status*	<MCPD_status>
17		*has_id*	:plant_identifier
18		*has_genus*	“genus”
19		*has_species*	“species”
20		*has_taxon_id*	<NCBI_ID>
21	:plant_identifier	rdf:type	<Accession (germplasm ontology)>
22		dct:identifier	“accession_ID”
23		*is_stored_in*	<*database*>

6 http://wiki.plantontology.org/index.php/Plant_Stress_Ontology

7 https://bioportal.bioontology.org/ontologies/PTO

8 http://www.cropontology.org/ontology/SP/Solanaceae%20Phenotype%20Ontology

Model structure. The developed model is presented in terms of semantic triples. The passport information is given with the MCPD, and AGRO is indicated for the experiment. Ontologies for the domain of plant breeding (Trait Ontology, Plant Ontology) can be used across crops, and supplemented by other crop-specific ontologies (Solanaceae Phenotype Ontology). Widely-used ontologies (prefixes rdf, geo, dct – dublin core terms, RO - Relations Ontology, SIO - Semantic science Integrated Ontology) could be used for generic terms. However, many predicates appropriate for this use, have not been defined in published ontologies, hence the need for placeholders.

The core of this model is the Experiment. For the sample queries presented in section (3.2) this is not necessary, but it allows the model to be easily extended with, for example, treatments and management. For now, the date of the experiment, its location and title are attached to it. Each experiment has a set of observations, each of which is made on a specific, physical plant or accession. The observation consists of one phenotypic variable, and the value for the trait being observed. Each plant has a local identifier for the specific experiment. It may possess more broadly used identifiers, through its link to, for example, an accession. To cover the cases of crosses, an ancestral plant is introduced, as a descendant of the physical entity in the experiment. In the case where an accession identifier is not used, another property (like genotype) might be used instead. Further MCPD attributes (such as common crop name, institute information, addresses and coordinates, taxon authorities etc. would also be specified here, including the holding institute from which this identifier originated.

Placeholders. Many terms in the table do not refer to a specific ontology or vocabulary. Especially for predicates, the lack of suitable terms is a hindrance for good semantics. Even in the prominently featured MCPD, those are lacking, and do not give any means (properties) to connect an entity with, for example, its biological status - though they do contain the relevant classes. Issues like this may already be subject of attention, but have not yet been resolved. Additionally, it is noted that these terms come from a variety of ontologies, the terms of which are not defined to be compatible. Therefore, it is imperative that, for such an example to be semantically correct, this needs to be amended, and constraints need to be more appropriately defined.

3.2 The demonstrator

The demonstrator itself, (Fig. 1) uses the above semantic model, and data available from the EU-SOL database (https://www.eu-sol.wur.nl/). The example questions were formulated in collaboration with PGR users and PGR providers, and they all require to query the MCPD for the “accessions” and the “crops”, as well as location data. These questions were hard-coded in the demonstrator (Fig. 1). As a user, one has to select the relevant query and specify its parameters (like the phenotype to search for, the desired country of origin, biological status, accession name). As we focused on tomato, the Solanaceae Phenotype Ontology was used. The options for each parameter are queried on the fly and displayed in a drop-down list. Accordingly, a SPARQL query is formulated, and run against the provided sources.

Limitations. The demonstrator currently does not search for relevant datasets across FAIR data points by itself. Instead, it retrieves the data from hard-coded resources, in the form of RDF, formatted according to the FDP definition. However, the demonstrator will be adapted to consume data from distributed resources once the relevant FDP's, formatted according the heretofore mentioned data model, are coming online; which also would enhance the possibility to query these resources directly by machines (e.g. via the SPARQL query language). The demonstrator is online at https://www.plantbreeding.wur.nl/ld-demonstrator/.

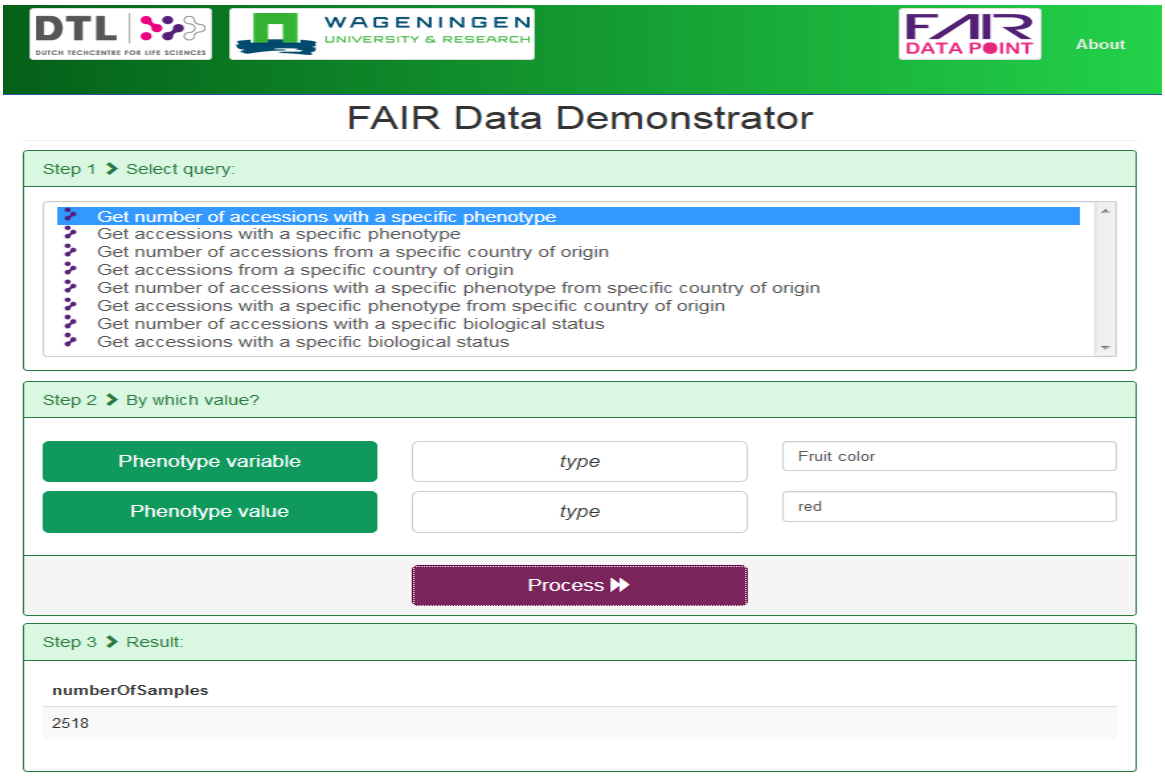

Fig. 1. The demonstrator interface: the user selects a question (highlighted), a phenotypic variable ("fruit color"), as well as a value for it ("red")

4 Discussion

Outcome. This demonstrator was developed to showcase how FAIR data infrastructures contribute to the sharing of PGR data. The result is a responsive graphical interface, answering predetermined questions but allowing more flexible querying via SPARQL queries directly. The value of this effort does not come from any novel questions posed, but from the distributed nature of the available PGR resources. Work on the semantic data model brought up some significant gaps that currently exist in the semantics that should be addressed in the future, such as the

placeholders in Table 1. In spite of those, the approach followed is a good example of such a process, highlighting the reusability of existing components.

Modeling pitfalls. The most demanding part is the construction of a semantic model. Lessons learned include: one should not deviate from designing a model reflecting the "real world" conditions, in favor of modelling for a specific dataset or database. This is to reaffirm that a specific database or entity-relationship diagram (ERD) is easily translatable into semantic triples, but does not necessarily lend itself to a schema that is intended to accommodate data from different providers.

Future work. In the future, the demonstrator will be extended to include more domain-relevant queries and implementation of the FDP infrastructure by PGR providers. As plants are "unable to move", we plan to explore the potential of geo-aware queries. However, the main challenge will be in the full integration with other data sources, such as weather or especially ~omics databases. Only then could big data technologies help to revolutionize plant breeding and have a significant impact on the world's food and nutrient security.

Acknowledgements. This work was supported by Luiz Bonino and Kees Burger (Dutch Techcenter for Life Sciences), as well as Marco Roos (Leiden University Medical Center) and Patrick Hendrickx (Wageningen University and Research). Their advice, technical expertise and source material contributions are highly appreciated.

References

1. Gray, E., Jennings, W., Farrall, S. & Hay, C. Small Big Data: Using multiple data-sets to explore unfolding social and economic change. Big Data & Society January-Ju, 1–6 (2015).
2. Spindel, J. E. & McCouch, S. R. When more is better: how data sharing would accelerate genomic selection of crop plants. New Phytologist (2016). doi:10.1111/nph.14174
3. FAO/Bioversity, Multi-crop passport descriptors v.2. (2012) [Online], Available: http://www.bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v21-mcpd-v21/, [Accessed: 8-Sept-2017].
4. The Editors. FAIR principles for data stewardship. Nature Genetics 48, 343–343 (2016).
5. Roos, M., Wilkinson, M., Kaliyaperumal, R., Thompson, M., Carta, C., Cornet, R. and da Silva Santos, L.O.B., Registries of domain-relevant semantic reference models help bootstrap interoperability in domains with fragmented data resources, Proceedings of the 9th International Conference Semantic Web Applications and Tools for Life Sciences, Amsterdam, The Netherlands, December 5-8, 2016.
6. Khoury, C., Laliberté, B. & Guarino, L., Genetic Resources and Crop Evolution (2010), Volume 57, Issue 4, pp 625–639 https://doi.org/10.1007/s10722-010-9534-z
7. Mackay M. C. (2011), Surfing the Genepool: the Effective and Efficient Use of Plant Genetic Resources, Doctoral thesis No. 2011:90, Acta Universitatis Agriculturae Sueciae, Swedish University of Agricultural Sciences, Faculty of Landscape Planning, Horticulture and Agricultural Science, Alnarp, Sweden.