ISSN 1479-2621
Plant Genetic Resources: Characterization and Utilization (2015) **13**(1); 90–93
doi:10.1017/S1479262114000689

Short Communication

Genebanks and genomics: how to interconnect data from both communities?

Richard Finkers[1]*, Pierre-Yves Chibon[1†], Rob van Treuren[2], Richard Visser[1] and Theo van Hintum[2]

[1]*Wageningen UR Plant Breeding, Wageningen University and Research Centre, Wageningen, The Netherlands and* [2]*Centre for Genetic Resources, Wageningen University and Research Centre, Wageningen, The Netherlands*

Received 26 February 2014; Accepted 3 May 2014 – First published online 27 May 2014

Abstract

Genebanks are important suppliers of genetic resources to the genomics research community, and access to the resulting information will allow traditional genebank users to better select genetic material for their breeding and scientific programmes. We discuss herein a possible solution to interconnect these data automatically based on semantic web technology.

Keywords: data integration; genebank community; genomics community; semantic web

Introduction

Genebanks generally store their data in straightforward custom-build databases. These databases always contain passport data, defining the identity and origin of the material. Other information stored may include characterization and evaluation (C&E) data, describing the phenotype of the accessions, logistic/management data about where the material is physically stored, germination test results, distribution data, etc. The computerized documentation usually stops there. The information should be made available through a website, enabling the users to select and request accessions.

Due to advances in molecular biology, which resulted in technology allowing high-throughput sequencing, and the large-scale application thereof, it has become apparent that genebank data and this type of genomics data will have to be connected. Genebanks are important suppliers of genetic resources to the genomics research community, and access to the resulting information will allow traditional genebank users to better select genetic material for their breeding and scientific programmes (McCouch *et al*., 2012; van Treuren and van Hintum, 2014).

A thus far unresolved question relates to the interaction; how will genebanks use and provide access to genomics data and how will genomics information resources give access to genebank data and material? Herein, we discuss a possible solution that allows making curated passport and C&E data online available on the genebank side and sequences and their annotations on the genomics data provider side, including public resources such as NCBI databases (Anguita *et al*., 2013) and UniProt (Redaschi and Uniprot Consortium, 2009). This approach will allow to interconnect these data; a genebank can provide access to the annotated sequences or search for allelic variants for specific genomic regions within its genebank material; a genomic database can give access to the details about the origin of the accessions, or the phenotypic data available at the genebank. This interconnection is based on semantic web technology, an established framework of the World

* Corresponding author. E-mail: richard.finkers@wur.nl
† Present address: Red Hat, Inc., Paris, France.

Wide Web Consortium (W3C, 2013). It is fairly simple to implement in a well-organized genebank.

The technology

The principle of the semantic web technology is that information is presented in a way that allows computers to find it, interpret it and link it to other information. Or in the words of the W3C: 'The Semantic Web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.' Implementation procedures are relatively simple and involve two steps: (1) the standardized definition of terms, i.e. an ontology; (2) the use of these terms when presenting information on the Internet.

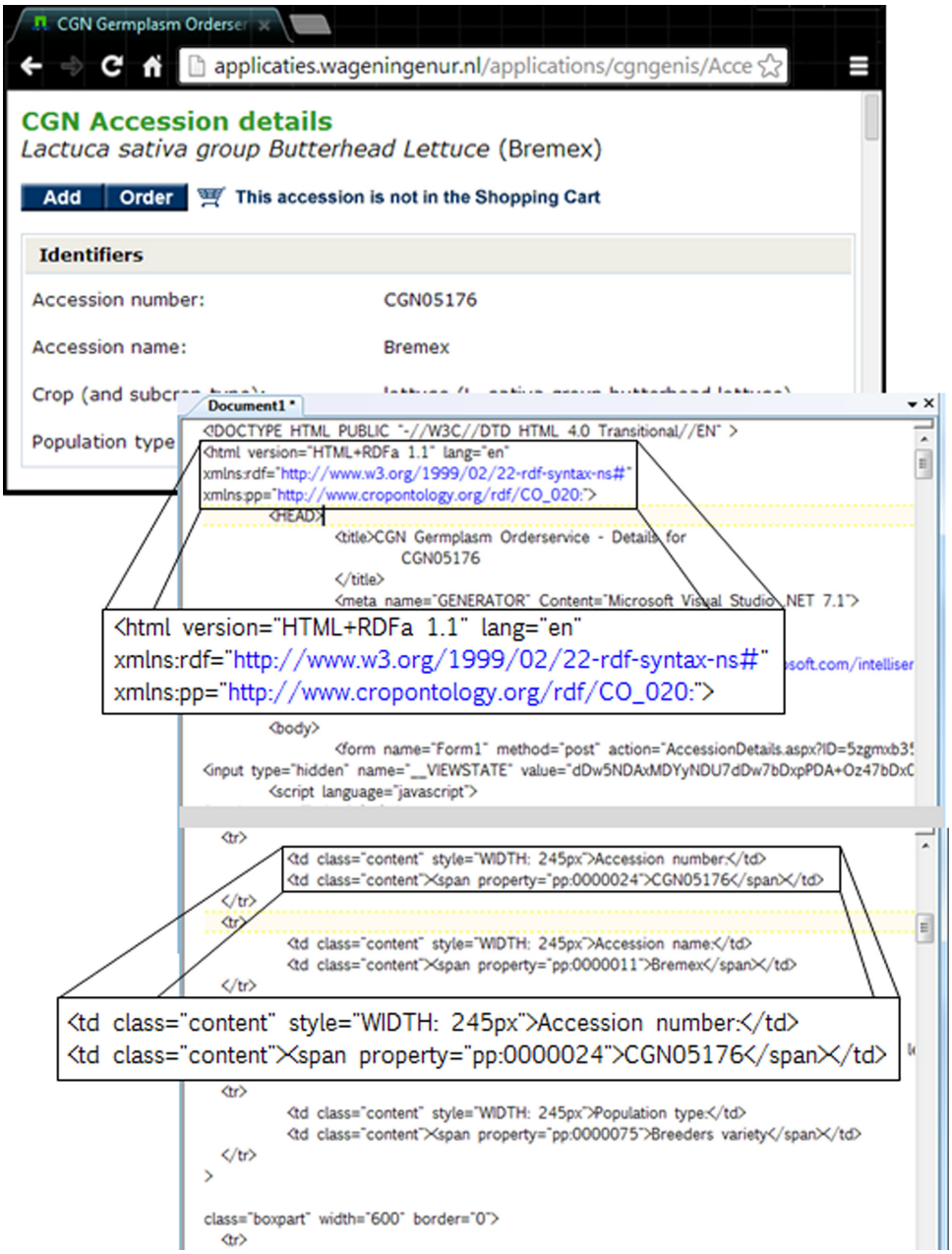

Fig. 1. Embedding of ontologies within a genebank accession result page. Each page starts with a reference to the ontology used in the tags, and descriptors from these ontologies are embedded in the webpage using the tags. These tags are no more than the codes of the descriptors from the ontology embedded in the proper formatting. In this example, we highlight the reference to the FAO/Bioversity Multi-Crop Passport Descriptor (CO_020) and the property 'donor accession number' (0000024). A semantic web-aware browser can interpret these tags automatically.

https://doi.org/10.1017/S1479262114000689 Published online by Cambridge University Press

The Crop Ontology (GCP, 2013; Matteis *et al.*, 2013), giving access to a variety of crop-specific and generic germplasm ontologies, proved suitable for genebank purposes.

The easiest method to add semantics is via embedding the information in the webpage. Each page starts with a reference to the ontologies used in the tags (the XML namespaces, for example xmlns:pp = http://www.cropontology.org/rdf/CO_020:). Descriptors from this ontology are embedded in the webpage using the tags (Pohorec *et al.*, 2013; Fig. 1). These tags are no more than the codes of the descriptors in the ontology embedded in the proper formatting. For example, the tag CGN and CGN05176 indicates that it concerns accession CGN05176 maintained by the Centre for Genetic Resources, The Netherlands (CGN). This method is suitable for making simple data available. For more complex data types, development of an application programming interface is the preferred method (W3C, 2008).

The outlined approach only requires the mentioned elements and an aggregator that knows where to find the relevant information, preferably via a centralized service that allows registration of information sources. The aggregator can be developed as a stand-alone tool or integrated in a website.

Use cases

To illustrate the potential of the semantic web technology, two use cases will be presented. The first involves a proof-of-principle for combining C&E data from a breeding database and a genebank database; the second involves the concept for a genebank website to present the allelic variants of a given gene available in its collection.

Prototype: combining C&E data

Information from a breeding database (containing ~omics data for a specific crop) can be combined with information from genebank records automatically by an aggregator. A prototype aggregator (SemGem, 2013) was built to integrate data from the EU-SOL BreeDB (2013) with those of the CGN. The EU-SOL record is retrieved and embedded Crop Ontology tags (passport and *Solanaceae* phenotype ontology) are extracted, including the reference to the genebank record. Information of this genebank record is extracted and presented in a table. This prototype shows the feasibility of such integration, provided that both data sources use semantic annotations. A more advanced integration of information is outlined in the next example.

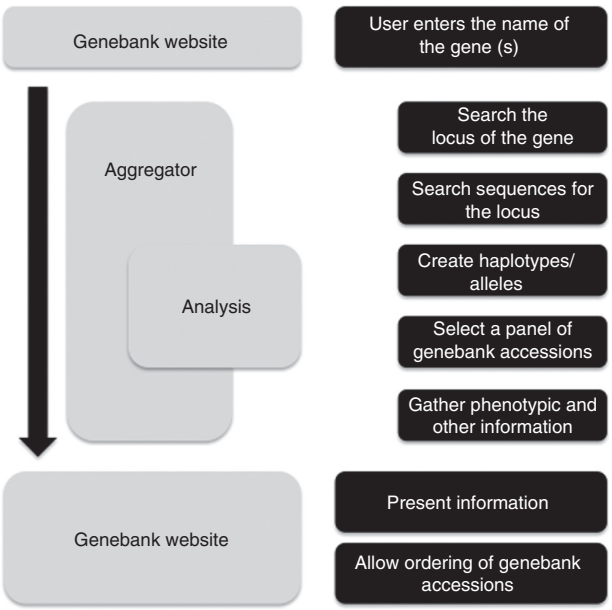

Fig. 2. Steps involved in the conceptual interface for presenting a user with a panel of accessions representing the complete allelic diversity of a given gene.

Concept: allelic variants in the genebank

A genebank website allows the user to enter the name of a gene, and the interface returns a list of accessions with different allelic variants for this gene, including allele name, known function and information about the phenotype (Fig. 2). Following this step, resources providing allelic variant data regarding that gene generated in one of the many genome (re-)sequencing initiatives (e.g. Lam *et al.*, 2010; Finkers *et al.*, 2012; Xu *et al.*, 2012) are aggregated from all available up-to-date resources on the Internet. Based on the resulting information, haplotypes are reconstructed for the queried gene and the germplasm will be organized in the identified haplotype groups. For each haplotype group, C&E data will be retrieved from the genebank database and presented in combination with the haplotype groups. This information should assist the user in selecting a germplasm panel suitable for his/her purposes.

Conclusions

To exploit the information generated by the genomics community, genebanks should link their data to those in genomics databases, rather than aiming to incorporate these information resources. The enabling technology to

https://doi.org/10.1017/S1479262114000689 Published online by Cambridge University Press

do so is available. Ontologies, defining the common terms and standards, play a crucial role in this effort and need to be adopted and improved by the genebank community. Implementation of the 'allelic variant' concept will be the next step towards interoperable genebanks.

References

Anguita A, García-Remesal M, de la Iglesia D and Maojo V (2013) NCBI2RDF: Enabling full RDF-based access to NCBI databases. *Biomed Research International* 2013: 983805.

EU-SOL BreeDB (2013) Available at https://www.eu-sol.wur.nl (accessed November 8 2013).

Finkers R, Smit S, Peters S, Schijlen E, van Heusden AW and Zhang G (2012) 150 Tomato genome (re-) sequencing project. In: *9th Solanaceae Conference*, Neuchâtel, Switzerland.

GCP (2013) Crop ontology curation tool, a service of the GCP integrated breeding platform. Available at http://www.cropontology.org (accessed November 8 2013).

Lam HM, Xu X, Liu X, Chen W, Yang G, Wong FL, Li MW, He W, Qin N, Wang B, Li J, Jian M, Wang J, Shao G, Wang J, Sun SSM and Zhang G (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. *Nature Genetics* 42: 1053–1059.

Matteis L, Chibon P-Y, Espinosa H, Skofic M, Finkers R, Bruskiewich R, Hyman G and Arnaud E (2013) Crop ontology: vocabulary for crop-related concepts. In: *Proceedings of the First International Workshop on Semantics for Biodiversity*, vol. 979, 27 May 2013, Montpellier, France.

McCouch SR, McNally KL, Wang W and Sackville Hamilton R (2012) Genomics of gene banks: a case study in rice. *American Journal of Botany* 99: 407–423.

Pohorec S, Zorman M and Kokol P (2013) Analysis of approaches to structured data on the web. *Computer Standards & Interfaces* 36: 256–262.

Redaschi N and Uniprot Consortium (2009) UniProt in RDF: tackling data integration and distributed annotation with the semantic web. *Nature Precedings* 272: 21784–21792.

SemGem (2013) Semantic genebank integration, prototype software for the integration of data of a breeding database with those of a genebank. Available at https://github.com/PBR/SemGem (accessed January 2014).

van Treuren R and van Hintum TJL (2014) Next-generation genebanking: plant genetic resources management and utilization in the sequencing era. *Plant Genetic Resources Characterization and Utilization* 12: 298–307.

W3C (2008) SPARQL query language for RDF. Available at http://www.w3.org/TR/rdf-sparql-query (accessed January 7 2014).

W3C (2013) Semantic web. Available at http://www.w3.org/standards/semanticweb (accessed 30 October 2013).

Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang J and Wang W (2012) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nature Biotechnology* 30: 105–111.

https://doi.org/10.1017/S1479262114000689 Published online by Cambridge University Press