# Final Report
# RDFENO

21 October 2017
Pommier Cyril Cyril.pommier@inra.fr
Ester Dzale Yeumo Kabore esther.dzale-yeumo@inra.fr
Michael Alaux Michael.alaux@inra.fr

## Executive Summary

One of the outputs of the RDA Wheat Data Interoperability (WDI) working group is a set of guidelines for data sharing along with use cases. The guidelines include phenotyping metadata recommendations like Minimal Information About Plant Phenotyping (MIAPPE) or ontologies for data annotation like the Crop Ontology. The objective of the current project is to implement some of those recommendations for data publication and to share a rich public dataset as RDF in a dedicated SPARQL endpoint. We want also to improve recommendations for this specific use case.
The targeted wheat dataset is the public INRA Breeding Network, available in GnpIS (http://dx.doi.org/10.15454/1.4489666216568333E12 ), a platform involved in the Elixir European project. It includes experimentations in six locations over fifteen years

## Objectives

This project objective is to demonstrate the feasibility of generating a dataset in the RDF format following the MIAPPE specification from a rich dataset about winter wheat phenotypic data. It has been generated by a French experimental network and includes observations for different characters (agronomic, quality, disease, phenology,…) on more than 10 experimental locations during more than 15 years and for more than 1700 winter wheat varieties. The success of this operation must be demonstrated by publishing the resulting RDF in a queriable SQPARQL system, publicly available through a web interface. For this work to be useful for the international community, it must also be reproducible and adaptable to other datasets. Therefore, guidelines and toolset must be published by the RDA IGAD group on the WheatIS data standards site (http://wheatis.org/DataStandards.php ) .

## Initial State

Before the beginning of the RDFENO project, the RDA IGAD had published recommendations for data publication, including in particular a rich section on plant phenotyping data. Indeed, while some datatype, such as genomic or genetic variability are quite homogenous among scientific communities, phenotyping is highly heterogeneous and therefore difficult to standardize. Three international communities, represented in the IGAD group, are involved on this particular question: Elixir, the European bioinformatics infrastructure, EPPN & Emphasis, the European (and beyond) infrastructure for Plant Phenotyping and the Consultative Group on International Agricultural Research (CGIAR). They produced in particular two standards, the Minimum Information About Plant Phenotyping Experimentation (MIAPPE, www.miappe.org ) and the Crop Ontology (www.cropontology.org ).
The crop ontology provides vocabularies and ontologies to describe and annotate plant experiment variables and parameters plus a formalism and a framework for their

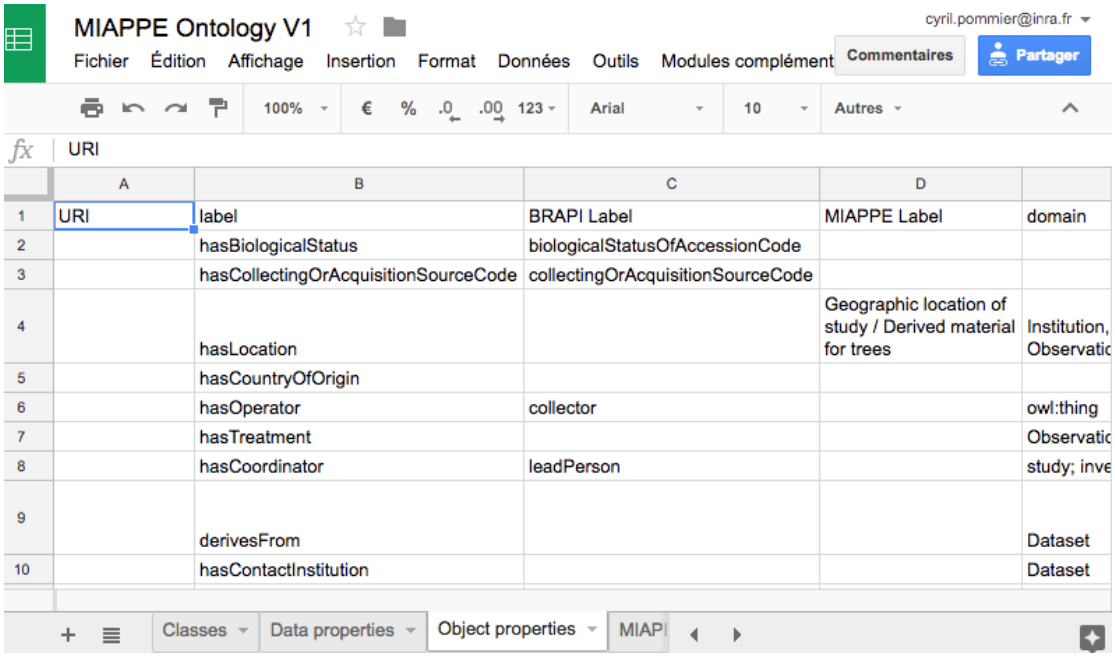
maintenance. They have proved their usefulness and their adoption is growing. Therefore, they are not within the scope of the present project.
MIAPPE is a list of minimum information to allow FAIR (Findable, Accessible, Interoperable and Reusable) data publication. Its adoption is growing. It has been implemented in some databases, in an archive format (ISA Tab) and in a web service specification (Breeding API). The next step is to add a further implementation as RDF.

# Project Outcomes

## Module 1

The objective is to build and document the process for performing a RDA-WDI compliant transformation of phenotyping data to a semantically interoperable RDF format, including methodology, tools, and documentations. This process will be detailed in a public online document.
The generation of a MIAPPE compatible RDF dataset has capitalize on the MIAPPE efforts through two main tasks: (i) build a MIAPPE domain ontology (MIAPPE RDF) and (ii) develop a toolset to generate RDF from a Breeding API web service endpoint.

The MIAPPE community got involved on the first task in a dedicated Workshop in Lisbon which was organized thanks to hosting by ELIXIR-PT and RDA RDFENO funding to invite CGIAR and Emphasis experts. The first version of the result ontology will be maintained in a dedicated github repository: https://github.com/MIAPPE/MIAPPE-ontology . It has been constructed through a shared Google sheet and will be converted to owl by the end of 2017.




| | A | B | C | D | |
|---|---|---|---|---|---|
| 1 | URI | label | BRAPI Label | MIAPPE Label | domain |
| 2 | | hasBiologicalStatus | biologicalStatusOfAccessionCode | | |
| 3 | | hasCollectingOrAcquisitionSourceCode | collectingOrAcquisitionSourceCode | | |
| 4 | | hasLocation | | Geographic location of study / Derived material for trees | Institution, Observatic |
| 5 | | hasCountryOfOrigin | | | |
| 6 | | hasOperator | collector | | owl:thing |
| 7 | | hasTreatment | | | Observatic |
| 8 | | hasCoordinator | leadPerson | | study; inve |
| 9 | | derivesFrom | | | Dataset |
| 10 | | hasContactInstitution | | | Dataset |



The second task has been developed at INRA-URGI thanks to RDA RDFENO budget which allowed to invest two person months of non-permanent engineer.

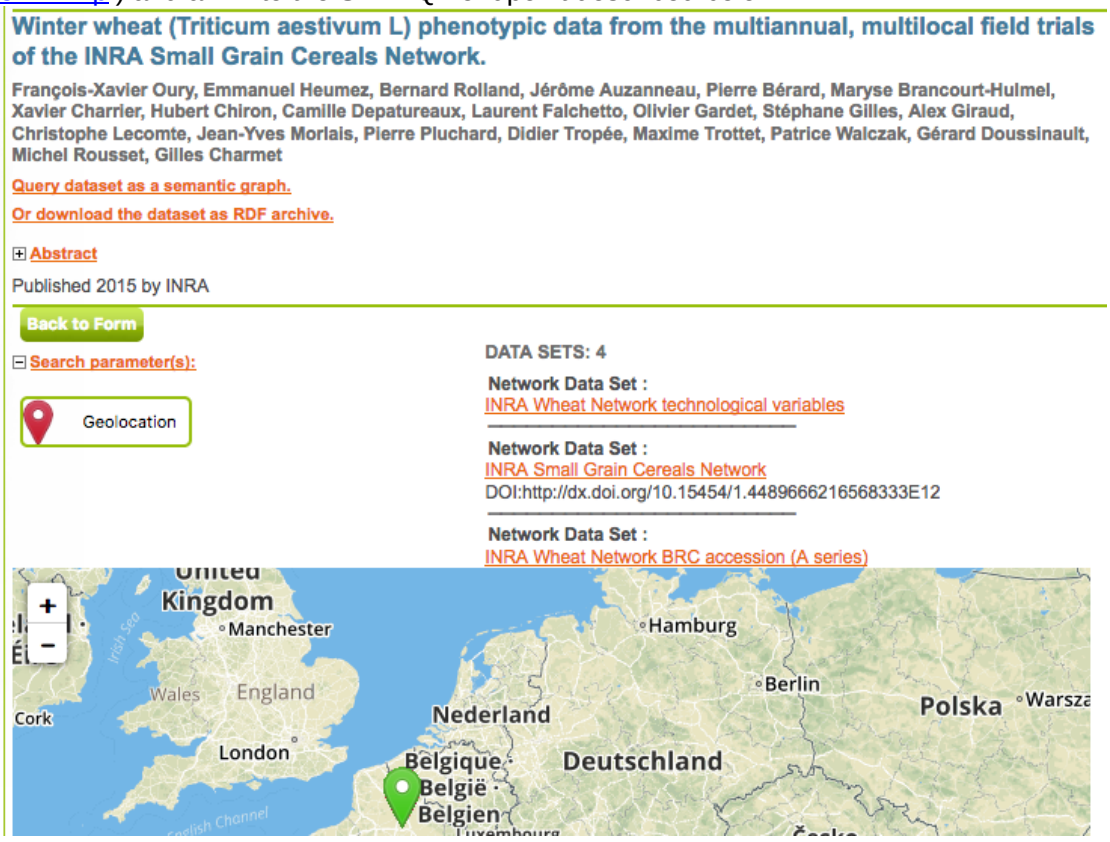
We have demonstrated that, even with a well curated dataset, there are some elements that need further improvement. For instance, some resources within the dataset have already a good identification, like for instance the plant individual or accessions which can have Permanent Unique Identifier (PUI) like DOI or URI, while other, like the treatments, lack such identification. We have therefore distinguished between elements that need a very large interoperability beyond a given dataset, and therefore that need a PUI, and the elements that need only PUI valid within the dataset.

Secondly, we have developed a BrAPI to RDF toolset which relies on JSON-LD and several transformations. The resulting RDF relies for now on its own data model which will be replaced latter by MIAPPE RDF (once version one is sufficiently stable).

This procedure has been documented and published on the WheatIS standards page: http://ist.blogs.inra.fr/wdi/phenotypes-as-rdf/. It will be improved and enriched based on the community feedback.

## Module 2

The objective is to publish the actual phenotyping dataset as a set of files following the RDF standard.
All the published data are available through the data set DOI (http://dx.doi.org/10.15454/1.4489666216568333E12) web page which is used as a long-term data access page and provides a link to the SPARQL endpoint and the dataset download link. Following this URL in a web browser gives access to the full description which include a link to a zip archive containing all the RDF files (https://urgi.versailles.inra.fr/files/ephesis/wheat_network/10.15454_1.4489666216568333E12.zip ) and a link to the SPARQL endpoint described below.



The access to this zip archive could be modified later with content negotiation through datacite DOI server to provide a more machine friendly access to the data. The following command illustrate this idea:

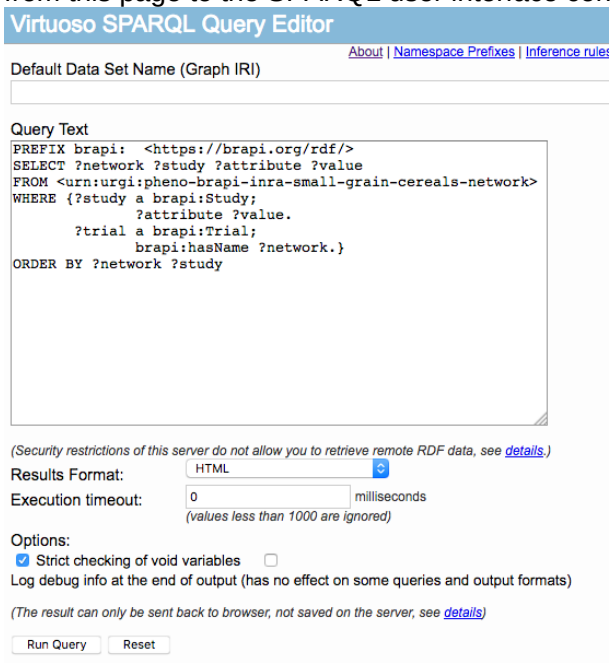
curl -LH "Accept: text/turtle" https://doi.org/10.15454/1.4489666216568333E12

### Module 3

The objective is to deploy a server to share this dataset through a SPARQL endpoint.
A Virtuoso SPARQL server (https://virtuoso.openlinksw.com/ ) has been installed and deployed specifically for the RDFENO project. The RDF dataset has been inserted in it and made publicly available through the result of Module 4.

### Module 4

The objective is to deploy a web user interface to query the dataset using the SPARQL language alongside examples based on WDI use cases.
The Virtuoso instance provides a public web interface to allow easy querying which is available at the following address: https://urgi.versailles.inra.fr/sparql. The long-term availability of this link will be will be ensured from the dataset DOI landing page. The link from this page to the SPARQL user interface contains a query example.




Due to limitations of this query interface, it has not been possible to add more examples. This interface will be either improved or replaced by another technology, like www.agrold.org or http://yasr.yasgui.org.

### Module 5

The objective is to provide a final report on the executed tasks and the results.
The current report is the result for this module

## Dissemination Activities / Publications

Dissemination actions have started and more are planned. The WheatIS data standards page was used as a way to advertise about the result of this project given the exemplar dataset and also because wheat data standardization is the subject of a recent publication (http://dx.doi.org/10.12688/f1000research.12234.1). It has already been advertised in

Elixir Plant Community and will be further disseminated within this infrastructure. It will be published also on www.miappe.org . Finally, we plan to present this work during one of the workshops of the Plant and Animal Genome (PAG) 2018 at San Diego.

| # | Event | Date | Publication |
|---|---|---|---|
| 1 | Conference PAG 2018 | | |
| 2 | | | |
| 3 | | | |
| 4 | | | |

## Summary & Conclusions

This project has been a good opportunity to gather key actors of the plant phenotyping community data management around MIAPPE data publication. Thanks to this support from RDA, it has been possible to choose and produce a single domain ontology which structures and define data for three important infrastructures. It has been therefore possible to avoid building competing standard, and in this respect, the project is a success.

We have also produced a very rich plant phenotyping dataset with a reproducible, documented and published procedure. It will not only useful to demonstrate FAIR RDF data access, but it will also enable novel data valorization, exploration and analysis. Finally, a lot of effort has been made to ensure that the results of this project will not remain in a silo but will be applicable to other datasets.