Opening and Linking Agricultural Research Data[1]

Esther Dzalé Yeumo Kaboré
French National Institute for Agricultural Research
edzale@versailles.inra.fr

Devika Madalli
Indian Statistical Institute
devika@drtc.isibang.ac.in

Johannes Keizer
Food and Agriculture Office of the United Nations
johannes.keizer@fao.org

Abstract

A Research Data Alliance (RDA) Interest Group has formed around a community of scientists and researchers who wish to make agricultural data, information, and knowledge more accessible. The objective of the Agricultural Data Interest Group is to get together active representatives of the major international institutions that work on agricultural research and innovation worldwide, in order to address issues related to data that are important to the development of global agriculture. The Interest Group is moving toward this goal by advancing the formation of a Wheat Data Interoperability Working Group that will address diverse data problems of 'wheat data' and, within a time period of 18 months, provide a framework for wheat data interoperability. The framework will foster the adoption of common standards and vocabularies for wheat data management, and facilitate access, discovery, reuse, and integration of that data.

1 Introduction

Agricultural data tends to be complex owing to its origin from different purposes, task oriented systems, organizations and projects. One of the main topics for discussion is the infrastructural requirement for scientific agricultural data and its management in standard forms and methods. There are several worldwide data initiatives in the agricultural domain that need to be interoperable. For example, agricultural datasets may be information resources

[1] Copyright © 2014 Esther Dzalé Yeumo Kaboré, Devika Madalli and Johannes Keizer
D-Lib Magazine.Vol.20(1/2):January/February 2014

or numerical data regarding yield, market prices or other associated data such as soil analysis data. The challenge is being able to provide intuitive services to end users that span continents with widely varying requirements in terms of data-based services. There are various stakeholders, from data producers to consumers, and their needs have to be satisfied in terms of services and researched products.

In recent years, efforts to make agricultural data, information and knowledge more accessible have increased. For more than 5 years now, Global Forum on Agricultural Research[2] (GFAR), the Cooperative Group on International Agricultural Research[3] (CGIAR) and numerous other partners have been promoting the Coherence in Information for Agricultural Research for Development[4] (CIARD) movement to open up access to agricultural knowledge worldwide. Within the CIARD movement the Routemap to Information Nodes and Gateways[5] (RING) was established, which already contains 986 data sources. In April 2013 the G8 group organized an event[6] on open data in agriculture, with participation far broader than the G8 organizers. At that conference, Secretary Vilsack (Agriculture, US Government) said that "open data" means "available without restrictions" and "machine readable". Since then, responding to an initiative of the UK and US governments, another international alliance was formed under the title of "Global Open Data for Agriculture and Nutrition"[7] (GODAN).

2 RDA Interest Group on Agriculture Data

An Interest Group (IG) on agriculture data has been initiated under the Research Data Alliance[8] (RDA). The Agricultural Data Interest Group[9] will link up with the international undertakings described above and will try to integrate them into the interdisciplinary framework of RDA. The objective is to get an active representation of the major international institutions that work on agricultural research and innovation worldwide and address issues related to data that are important to the development of global agriculture. The group intends to bring together leaders and stakeholders in agricultural data initiatives to work towards feasible services in a technology-independent and standards-compliant platform for worldwide use by communities in agricultural sectors, and to represent all stakeholders producing, managing, aggregating, sharing and consuming data for agricultural research and innovation.

[2] http://www.egfar.org/ Accessed January 2014.
[3] http://www.cgiar.org/ Accessed January 2014.
[4] http://www.ciard.net/ Accessed January 2014.
[5] http://ring.ciard.net/ Accessed January 2014.
[6] https://sites.google.com/site/g8opendataconference/home Accessed January 2014.
[7] http://www.godan.info/ Accessed January 2014.
[8] https://rd-alliance.org/ Accessed January 2014.
[9] https://rd-alliance.org/ Accessed January 2014.

The IG has already begun to gather key information needed to support the development of infrastructure within a proposed Wheat Data Interoperability Working Group that will address diverse data problems of 'wheat data' and, within a time period of 18 months, provide a framework for wheat data interoperability. As a first practical step, the IG is preparing a Case Statement[10] for the Wheat Data Interoperability Working Group. Meetings to define the group have already been held.

3 Wheat Data Interoperability

Interoperability is a wide concept that encompasses the ability of organisations to work together towards mutually beneficial and commonly agreed goals. The Agricultural Data Interest Group is using the following definition from the European Interoperability Framework (EIF):

> *'An interoperability framework is an agreed approach to interoperability for organisations that wish to work together towards the joint delivery of public services. Within its scope of applicability, it specifies a set of common elements such as vocabulary, concepts, principles, policies, guidelines, recommendations, standards, specifications and practices.'*

The proposed Wheat Data Interoperability Working Group will aim to provide a common framework for describing, representing linking and publishing wheat data with respect to open standards. Such a framework will promote and sustain wheat data sharing, reusability and operability. Specifying the wheat linked data framework will require answering many questions, including which (minimal) metadata to describe which type of data; which vocabularies/ontologies/formats; and which good practices.

Mainly based on the needs of the Wheat Initiative Information System (WheatIS) in terms of functionalities and data types, the Agricultural Research Interest Group and the proposed Working Group will identify relevant use cases in order to produce a "cookbook" on how to produce "wheat data" that are easily shareable, reusable and interoperable. To do so, the Working Group will:

- Run a survey of existing standards and recommendations (vocabularies, ontologies, formats). This survey will identify which standards are adopted in the wheat data managers community, which ones are missing and which ones can stand as references. (Interest Group)
- Run a survey to identify potential partners willing to share data in order to better understand the players who will need to be actively engage. Identify the main wheat data types, end-user categories, and case studies and provide standards harmonization,

[10] https://www.rd-alliance.org/groups/agricultural-data-interoperability-ig/wiki/wheat-data-interoperability-wg-charter-v3.html Accessed January 2014.

guidelines to describe, document, structure and interlink data, taking into account the diversity of data types. (Interest Group)

- Evaluate the interest of linked data technologies to improve usage and access to the information. (Interest Group)
- Identify and deploy relevant platforms to support the Wheat linked data framework. (proposed Working Group)

Based on a survey report performed in June 2012, the Working Group will focus on the following data types, by order of priority: SNP, Genomic annotations, Phenotypes, Genetic Maps, Physical Maps, Germplasm. Implementing the framework will help cultivate a wheat ecosystem with people familiar with interoperability, organisations ready to collaborate, and common tools and services.

The WheatIS[11] (Wheat Information System of the Global Wheat Initiative) will be provided with a linked data framework based on community-accepted standards (an RDA Wheat Interoperability proposed WG deliverable), which will ensure data analysis and data integration facilities. Such a framework is a great asset for the WheatIS to provide the analysis functions and other services expected by the researchers.

The cookbook will ensure that:

- Wheat data managers and data scientists will have a common and global framework to describe, document, and structure their data.
- Researchers, growers, breeders, and other data users will have seamless access, use, and reuse to a wide range of Wheat data. Data linking will also ease emergence of new data analyses and knowledge discovery methodologies.
- Other plant data managers and scientists — will have the benefit of a reusable data framework.
- Researchers working on other plants will be able to more easily access, reuse and link up Wheat data with their own data.

The cookbook might be adapted for other crops such as RICE, MAIZE which are also very important for food security.

There is a variety of wheat data sources with various types of data, formats and structures. Integrating these data is an important part of current research, specifically when answering questions such as "What genes and traits are relevant for understanding the impact of climate change on wheat plant productivity?". The wheat data interoperability framework should have an impact and be helpful by a) fostering the adoption of common standards and vocabularies for wheat data management, and b) facilitating access, discovery, reuse and integration of wheat data.

[11] http://www.wheatinitiative.org/research/wis Accessed January 2014.

4 Additional Discussion Areas within the Agricultural Data Interest Group

Data Policies

Data policies is a generic concept which includes more specific ones, such as data sharing, data management, IPR. All these should be taken under consideration for the work of the Working Group, and related work from other RDA groups should be adapted. It was noted that different data sources (e.g., public/private, institutional, national, worldwide; IMF/World Bank, etc.) have different data policies, so they could be explored. This item will be discussed in detail during the next IG meeting during the 3rd RDA Plenary Meeting[12] in Dublin, Ireland in March 2014. It would be beneficial for the Agricultural Data IG to invite a member of the RDA IG working on data policies to share experiences at the next IG meeting.

Proposal for a Germplasm Working Group

A Germplasm Working Group should be formed by the Agricultural Data IG, and work should be done towards enhancing interoperability, achieving standards and enabling data sharing and exposure for the biodiversity community, along with defining common global germplasm descriptions. This request was raised during the previous biodiversity meeting in Rome from Bioversity International[13].

Support

Support will be required for individuals to participate in the activities of the Agricultural Data IG, such as actual work, dissemination, etc. Support may be channelled either from related projects (such as agINFRA[14] in the case of Germplasm), related initiatives (e.g., Wheat Initiative) and possibly new projects funded under the Horizon 2020 programme or other sources. Since participation of the group members is voluntary, only a little time can currently be allocated to the activities of the group.

Sustainability

Existing business models should be studied in order to ensure sustainability. The example given was that even open data may disappear if triple stores are not financially supported. People often associate open data with free data, but there are costs that are associated with the curation and storage of the data. A fully articulated business proposal which could show the effect of opening data and how the linking of data benefits agriculture has not been generated yet. The advantages of opening databases for currently private systems should be highlighted.

Data sources

It is important to identify and harness data sources and aggregators that already have useful resources for the agricultural domain. Among the main objectives of the RDA Agricultural Data IG is fostering data sharing, data management and data interoperability. In this context,

[12] https://rd-alliance.org/rda-third-plenary-meeting.html Accessed January 2014.

[13] http://www.bioversityinternational.org/ Accessed January 2014

[14] http://aginfra.eu/ Accessed January 2014.

data sources such as Dataverse should be considered, since they have already taken steps towards linking and aggregating data and metadata in the agricultural sector.

As a next important step, stakeholder involvement in the Agricultural Data IG should be enlarged. The goal is not to secure a comprehensive coverage of all institutes and organizations, but rather representative coverage of practitioners who will able to work with, and for, the community.