



Improving access to research data

Anne-Françoise Adam Blondon

► To cite this version:

| Anne-Françoise Adam Blondon. Improving access to research data. 2017, 2 p. hal-02790301

HAL Id: hal-02790301

<https://hal.inrae.fr/hal-02790301>

Submitted on 5 Jun 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Improving access to research data

French National Institute for Research in Agriculture (INRA) explores the importance of ensuring research data is easy to access in the agriculture field

Facilitating the access to public data has always been an important endeavour in life science. The variety of data and sources has been growing even since the empirical sciences have arisen¹, questioning the communities on their best practices for data sharing and re-use. In addition to this ancient but still existing problem, the evolution of technologies used to generate data have added the challenges of their exponential volumes and speed of production. Nowadays, big data questions the agility and the sustainability of the data management systems that emerged in the 90's, essentially based on a backbone of international archives. The whole community is working towards a new model with the help of international organisations or consortia, such as the [Research Data Alliance](#) or the [Global open data for Agriculture and Nutrition](#). The necessity to enable a more distributed model for data management has emerged, to be combined with higher efforts for standardization at different levels and supported by an ecosystem of infrastructures, institutions, consortia of researchers and/or private companies, organisations².

Unit of Research in Genomic-Info (URGI)

The Unit of Research in Genomic-Info (URGI) at the French National Institute for Research in Agriculture (INRA) is a very active player in the plant community initiatives, towards enhancing access to data in plant biology. The



story started 15 years ago with the development of a centralised information system, [GnpIS](#)³, aiming at storing, integrating and giving access to different type of data on different crops and pathogens: genetic resources, genetic and physical maps, genomes and their annotations, polymorphisms, phenotypes, GWAS. More recently, URGI contributed to 2 initiatives, the TransPLANT infrastructure (FP7 EU program, n°283496) and the Wheat Information System (WheatIS) of the [Global Wheat Initiative](#), that allowed us to build the vision and a proof of concept for a federated information system for research in plant biology and breeding. The shared vision of the 2 projects was that information systems sharing the same global semantics should be able to program-

matically expose the description of their content to a unique web portal that could be used by researchers and breeders to search data. The technical bricks of this vision were built by the [TransPLANT](#) project and 2 demonstrations of the concept were implemented, one involving 9 European databases and the [other involving 12 international databases](#). The challenges that are now ahead are (i) to improve the technical system to facilitate its maintenance (e.g. make it possible for one node of the federation to update its software versions without affecting the whole system) and (ii) to allow increasingly user-friendly searches (e.g. be able to retrieve with "yield" query, anything in relation with yield but not necessarily associated to the yield world in the data set). These 2

challenges are currently endeavoured by the partners of the [ELIXIR](#) European infrastructure (Excelerate H2020 EU project, n°676559) and also specifically by the French node of ELIXIR for [some aspects](#) .

URGI and its ELIXIR partners have also been very much involved in the development of another technical brick that enabled improvements of such federation of databases: an international standard web service called the [Breeding API](#). The current version allows programmatic standard retrievals of data about the genetic material and of phenotyping data and is being implemented in relevant ELIXIR nodes.

In parallel to these “technical” proofs of concept, INRA has used the WheatIS project to develop with the wheat community of researchers and breeders a set of guidelines for making their data [findable, accessible, interoperable and re-usable](#) and meet the best the standards for open data. These guidelines are available on the WheatIS portal and have been implemented in a central file repository accessible from the same portal that was developed by URGi to complement the existing databases. These 2 resources help the data producers to describe the purpose and the content of a dataset in a way which is understandable by any new user. The European community of plant data managers, together with specialists of data standardisation and the data producers (e.g. partners of EMPHASIS, the European Infrastructure for Plant Phenotyping) have also worked very actively to develop

and improve a standard for phenotyping data that did not exist before: Minimal Information About Plant Phenotyping Experiment ([MIAPPE](#)). This standard was presented for discussion to the international community of plant scientists and adopted. Recently, a mechanism of governance has been set up at the initiative of ELIXIR to regulate its future evolutions. The plant science community is now working on transferring the tools and knowhow gained through the WheatIS and TransPLANT/ELIXIR experiences to all crop communities through actions of communication⁴ or training and implementation of good practices in biology driven projects.

An important part of the work achieved by plant biologists, however still relies on central archives and knowledge databases maintained for instance by EMBL or the NCBI. These resources are invaluable and ELIXIR is currently working on how to better identify additional core resources and what could be their sustainable business model.

Finally, it has to be stressed that the building of such an ecosystem of information systems to the benefit of the final user requires a lot of effort in community building, within and between several communities: the data producers across crops, the data managers, the developers and the specialists of standards, ontologies and semantics.

References

- 1 Strasser B. J., 2012. Data-driven sciences: From wonder cabinets to electronic databases. doi: 10.1016/j.shpsc.2011.10.009

- 2 Leonelli S et al., 2017. Data management and best practice for plant science. doi: 10.1038/nplants.2017.86

- 3 Steinbach D, et al., 2013. GnplS: an information system to integrate genetic and genomic data from plants and fungi. doi:10.1093/database/bat058

- 4 Adam-Blondon A-F, et al., 2016. Towards an open grapevine information system. doi:10.1038/hortres.2016.56



Anne-Françoise

Adam-Blondon

Director of Research

French National Institute for Research in Agriculture (INRA)

Tel: +33 1 30 83 37 49

anne-francoise.adam-blondon@inra.fr

<https://urgi.versailles.inra.fr/>

[www.twitter.com/inra_urgi](https://twitter.com/inra_urgi)