[Final version submitted to AGRIS on-line Papers in Economics and Informatics, http://online.agris.cz, published open access in Volume IV, Number 4, 2012, http://online.agris.cz/archive/2012/04/]

agINFRA: Building Blocks for a Data Infrastructure and Services to empower Agricultural Research Communities

G. Geser[1], Y. Jaques[2], N. Manouselis[3], V. Protonotarios[3], J. Keizer[2], M. Sicilia[4]

[1]Salzburg Research, Austria
[2] Food and Agriculture Organization of the United Nations, Italy
[3]Agro-Know Technologies, Greece
[4]University of Alcalá, Spain

Abstract

The agINFRA project aims to provide the agricultural research communities with e-infrastructure and services for open data access, sharing and re-use. This paper introduces the project's objectives and data principles, presents the data resources that are covered, and illustrates agINFRA services with examples from the area of agricultural statistics. Finally, it summarises how agricultural research institutions and other stakeholders can participate in, and benefit from, the project.

Key words

agricultural research, agricultural repositories, e-infrastructure, data infrastructure, open access, agricultural statistics

Introduction

The agINFRA project (www.aginfra.eu) is an EU-funded project under the 7th Framework Programme (FP7). The project develops data infrastructure and services for sharing results of agricultural research communities that are managed by international, national, institutional and subject-based repositories. The project involves technology and content partners from Europe, China, Ecuador and India, while one of the lead partners is the Food and Agriculture Organization (FAO) of the United Nations. Moreover the project goals are aligned with the strategic initiative Coherence in Information for Agricultural Research for Development (CIARD) that mobilizes and supports institutions in making agricultural research results more accessible globally (Pesce *et al*., 2011).

Many agricultural research organizations already have content repositories and portals that serve scientists, information officers as well as educators and extension workers, ranging from national/regional initiatives to global ones, like the Consultative Group on International Agricultural Research (CGIAR), one of the world's largest and most experienced global research organizations (Clark *et al*., 2011). What distinguishes agINFRA from these Web destinations can be illustrated with an analogy: The project will develop the infrastructure that helps passengers – information – get from place to place in an easy, secure and effective way. agINFRA does not develop the vehicles (e.g. cars and buses) that carry around the passengers. The main goal is to develop the infrastructure (road network, petrol stations etc.) that will allow others to transfer the passengers, i.e. exchange and share research information. Though agINFRA will also adapt and improve existing vehicles (e.g. Web content management systems and services) in order to show manufacturers how they can build better ones that will take advantage of the new infrastructure.

Materials and methods

e-Research infrastructures and services

Similarly to other infrastructures, agINFRA provides services that allow (research) communities to work together and the (data) economy to function. Research is becoming increasingly distributed, collaborative, ICT and information-intensive. As Hey and Hey (2006) note, e-science "is not a new scientific discipline in its own right: e-Science is shorthand for the set of tools and technologies required to support collaborative, networked science. The entire e-Science infrastructure is intended to empower scientists to do their research in faster, better and different ways."

Widely used definitions of e-research infrastructure have been outlined in the first roadmap of the European Strategy Forum on Research Infrastructures (ESFRI, 2006) and by the US National Science Foundation Cyberinfrastructure Panel (NSF, 2007). The latter defines e-research infrastructure as "cyberinfrastructure" that "integrates hardware for computing, data and networks, digitally enabled sensors, observatories and experimental facilities, and an interoperable suite of software and middleware services and tools. Investments in interdisciplinary teams and cyberinfrastructure professionals with expertise in algorithm development, system operations, and applications development are also essential to exploit the full power of cyberinfrastructure to create, disseminate, and preserve scientific data, information, and knowledge". The term cyberinfrastructure has been used in the context of research infrastructures related to life sciences, like the one proposed by the iPlant Collaborative, a project funded by the United States National Science Foundation (NSF) which created an innovative, comprehensive, and foundational cyberinfrastructure in support of plant biology research (Goff *et al*., 2011).

The European High-level Expert Group on Scientific Data (2010) understands that scientific data infrastructure "must be flexible but reliable, secure yet open, local and global, affordable yet high-performance". Also particularly important are principles of collaboration, trust and sharing of various resources in the networked research environment (content repositories, databases, software, networks, computing and other resources). The need for infrastructures supporting the researchers in their tasks has been identified by other scholars in many disciplines (Androulakis *et al*., 2009; Descher *et al*., 2009; Michener & Jones, 2012; Thessen & Patterson, 2011;). In this paper, we present the case of a data infrastructure for agricultural research sharing, explaining the rationale for its set up as well as the expected benefits for its users.

Open data principles and values

Overall, agINFRA means getting agricultural research data out of its silos. Helping open up and interlinking the data of existing and newly built repositories is a core activity of the project. The Linked Open Data principles, as suggested by Tim Berners-Lee (2009), and further elaborated by Bizer *et al*. (2009) are an important basis for this activity:

★	Make your data available on the web (whatever format), *but with an open licence, to be Open Data*.
★★	Make them available as machine-readable structured data (e.g. excel instead of image scan of a table).
★★★	As (2), but use non-proprietary formats (e.g. CSV instead of excel).
★★★★	All of the above, plus: Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff.
★★★★★	All the above, plus: Link your data to other people's data to provide context.

Table 1: Five stars of Linked Open Data (Berners-Lee, 2009)

In a broader perspective, agricultural research data that is shared through agINFRA will have to respect and serve the following desired values of scientific data:

Open – Data must be open and interlinked, not subject to barriers based on standard formats and, thereby, prevent data silos due to lack of interoperability and interrelatedness.

Meaningful – Data must be meaningful through explicit semantics, re-usable from available mature terminologies and ontologies that are exposed and interlinked through the Web.

Reliable – Data must be accessible with ensured provenance. Capability to express and trace the context of creation and re-use are important for building trust in research infrastructure services.

Actionable – Data must be actionable through services that empower research. The value of data is limited if researchers cannot act on it in the ways they need, using flexible and adaptable services.

Covering a broad range of data

Through agINFRA e-infrastructure and services many kinds of information relevant to agricultural sciences can be shared (Karampiperis *et al.*, 2012). A review of content domains of direct relevance to agricultural research identified some priority areas that serve as a starting point to build the agINFRA shared data space. Additional ones are also expected to be covered in the future, for example, cross-domain areas such as agro-biodiversity and agro-ecology (Benckiser & Schnell, 2006; Jarvis et al., 2007; Wezel *et al*., 2009).

At this stage, agINFRA is targeting the integration of five domains that cover both areas of specific research focus (e.g. agricultural economics) and areas where a particular type of information provides a platform for research activity in general (e.g. bibliographic resources). Currently the following domains are covered:

- **Bibliographic data on scientific and grey literature**, for example, FAO's AGRIS database (http://agris.fao.org) containing over 4 million bibliographic entries and records (Fogarolli *et al.*, 2011);

- **Digital learning and training resources**, for example, the Latin American Federation of Learning Object Repositories (LA FLOR - http://laflor.laclo.org) and the Organic.Edunet learning resources for organic agriculture and agroecology (www.organic-edunet.eu) (Dimitropoulos *et al.*, 2011);

- **Geospatial information systems** offering maps of land cover and soils, GIS datasets and other data with an agricultural or environmental theme (Aditya & Kraak, 2007), for example, the FAO GeoNetwork (www.fao.org/geonetwork/srv/en/main.home) and national resources such as the Italian Soil Information System (ISIS - http://aginfra-sg.ct.infn.it/isis);

- **Plant germplasm collections and genomics information**, for example, the Chinese Crop Germplasm Research Information System (CGRIS - http://icgr.caas.net.cn/cgris_english.html) and other national and international collections (e.g. European National Inventories of germplasm as shared through the EURISCO data catalogue); databases of DNA sequences and DNA barcodes;

- **Agricultural statistics**, for example, FAOSTAT (http://faostat.fao.org - over 3 million statistical entries, time-series data, etc.), other United Nations databases

and the World Bank open data catalogue (http://data.worldbank.org/data-catalog - providing access to over 8,000 indicators from World Bank datasets).

Examples of research data sharing in the area of agricultural statistics

Today agricultural statistical data are mainly available through major aggregated resources such as FAOSTAT and related United Nations' databases, the Organisation for Economic Co-operation and Development (OECD - www.oecd.org/statistics), World Bank (www.worldbank.org) and other international agencies as well as national economic data sources.

In comparison, the sharing of data collected by researchers working at universities and other research centers is rather limited. The main focus here is on providing access to research papers which, however, has reached considerable volumes. The research field avails of an increasing number of open access journals, many of which are covered by AgEcon Search (http://ageconsearch.umn.edu). AgEcon search is a free, open access repository of full-text scholarly literature from over 60 journals in agricultural and applied economics, including working papers, conference papers and journal articles.

A related European initiative has been the Network of European Economists Online (NEEO), coordinated by the Nereus Consortium (www.nereus4economics.info). The project developed the federated multilingual Economists Online portal (www.economistsonline.org/home) which draws on content repositories of 24 universities, including publications and datasets (Blake, 2009).

Probably the largest initiative is Research Papers in Economics (RePEc - http://repec.org), the collaborative effort of hundreds of volunteers in 75 countries to enhance the dissemination of research in economics and related sciences. RePEc provides a decentralized bibliographic database of working papers, journal articles, books / book chapters from over 1400 archives. In October 2012, RePEc comprised over 1.2 million records of 1500 journals and 3300 working paper series, of which 700,000 articles were available online. RePEc does not include research datasets, while AgEconSearch has a section on datasets that are freely available on the new AgEcon Search Dataverse (http://dvn.iq.harvard.edu/dvn/dv/AgEconSearch). However, since 2010 research groups only provided 5 datasets, which may illustrate the low level of preparedness for sharing of datasets in such ways.

Somewhat more advanced is the field of econometry. Econometrica (http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1468-0262), the journal of the Econometric Society provides a website of supplementary material "to enable replication of empirical and experimental work and other material related to papers that appear in the journal" (http://www.econometricsociety.org/suppmatlist.asp). Since Volume 72 (2004), over 230 papers with such supplemental material have been

published, however, only few papers in this journal relate to topics of agricultural economics.

The main question is: How can research institutes in the area of agricultural statistics on production and trade share and interlink their content and data more effectively?

Let us consider how an institute can open up its census, survey or time-series data by making them accessible to users through agINFRA-facilitated tools and services. As an example we use a Regional Fishery Body, the Secretariat of the Pacific Community (SPC - www.spc.int). SPC is a regional fishery body that monitors fish stocks in the South Pacific Ocean. It publishes yearly assessments of the fish stocks in its area of competence.

The SPC has already begun improving the dissemination of its data by participating in the Fishery Resources Monitoring System (FIRMS - http://firms.fao.org), a global network of regional fishery bodies sharing their assessments according to a common format. But they would like to go further, and make the survey data underpinning their assessments available as a global public good, thus allowing others to use it and making their assessments more transparent as well.

Step 1: Registration

The SPC data manager contacts the agINFRA consortium and makes a request to become a data provider. Once approved they are invited to register with the CIARD RING (http://ring.ciard.net), a global registry of agricultural data providers, datasets and services.

In the registry the data manager puts in not just the institution's contact details, but also describes the species capture production datasets that SPC would like to expose. To her delight, she finds that the ASFIS species classification (www.fao.org/fishery/collection/asfis/en) that SPC uses is already listed as an available dataset dimension in the agINFRA linked open data service that the CIARD RING accesses when users are defining their datasets. This simplifies the process as she does not have to describe or upload the scheme.

As SPC has no web service interface for the access of their statistics (one of the reasons they want to use agINFRA services) the data manager does not describe the service.

Step 2: Extract, Transform and Load (ETL)

The agINFRA ETL process allows the data manager to upload one CSV file per year of data, each containing two dimensions, species and area. These are automatically converted into RDF (Resource Description Framework) data cube format and stored in agINFRA's powerful triple store.

Step 3: Generation of multiple formats

Additional agINFRA transformation methods make the dataset available in several formats, including Statistical Data and Metadata Exchange (SDMX) and Google's Dataset Publishing Language (DSPL). SDMX defines representations of statistical data and respective metadata annotations, not only for single data items but also for full data sets (Gottron *et al.*, 2011). The dataset is also indexed for efficient searching across the infrastructure both internally and externally through an open search API. Finally the CIARD RING data is updated making users aware that this new dataset is available.

Step 4: Attachment of data to relevant research publications

The data manager receives a unique resolvable URI for each dataset. She is now able to attach these URIs to the current year's fish stock assessments, thus linking the documentary assessment to the raw research data on which the conclusions are based. She also uploads metadata for the documentary assessments into the infrastructure so that the documents can be searched and discovered together with the datasets.

Step 5: Recommending statistics related to other information resources

Users of the agINFRA recommender widget will automatically find the results from these time-series data appearing in their web sites. Mashups using the statistics widget will automatically get tables of statistics generated when their pages match the dimensions attached to the statistical data.

Results and Discussion

How agricultural research institutions and other stakeholders can participate and benefit

agINFRA is designed as an open and collaborative initiative. Therefore it offers a number of ways for stakeholders to engage in the agINFRA ecosystem of infrastructure and service developers, repositories, research organizations and educational institutions. The degree of involvement is decided by each participant according to the principle "the more you contribute, the more you can get back". Some of the key benefits for participants and contributors include:

Opening up research results (open science)

There is a wealth of raw, processed, analyzed and published agricultural science data that is collected and stored every day. Finding the way to make them accessible to the wider community will ensure that the research efforts are recognized and acknowledged. Provision of advanced tools and services will allow research organizations better organize, publish and interlink information about their content and data collections. Opening up theses collections to the international scientific

community will create more awareness of the research output and stimulate new collaborations.

Promoting data exchange

agINFRA's viability is tightly connected to the community of institutions and research groups that share through it new agricultural data sources and collections. Registering a collection as an agINFRA data source and publishing metadata for the resources in the collection ensures that they become part of a global pool of agricultural research results. Thereby research groups and individual scientists and educators will gain access to more relevant information for their work, also including other resource types than research papers and other documents.

Finding and re-using data

Agricultural research data of various types and formats will be made available by the agINFRA (meta)data pool. Different access protocols and formats are being put in place to allow this data to become searchable and consumable. Open search APIs, access protocols like OAI-PMH, and other types of add-on components and plug-ins will make it easier for existing systems to ingest data that reside in the agINFRA data pool. Simple solutions include harvesting the data of a particular type (e.g. bibliographic or economic information) and adding it to existing collections or search facilities.

Contributing software

agINFRA tools and services are being developed on an open-source code base, ensuring transparency, flexibility, and long-term viability of the software tools and applications that are being hosted, processed and empowered by the infrastructure. Software developers can use the agINFRA technical framework, components, add-on plug-ins and technical support for enhancing existing tools and services that are provided to agricultural researchers and data managers. Developers have the opportunity to participate in training events, plugfests and hackathons. These events will help gather their feedback and ideas and provide these back to the wider developer community.

Sharing Cloud and Grid infrastructure resources

An essential component of agINFRA is the availability of cloud and grid resources that various infrastructure partners are contributing. Access to the infrastructure is virtualized: clusters of servers are networked into an agINFRA Virtual Organization which is made available to the software tools and applications as a seamless infrastructure resource through a Scientific Gateway. Different middleware software components can be easily parameterized in order for a new infrastructure to contribute some of its cloud and grid resources to the agINFRA community.

Conclusions

The agINFRA project develops e-infrastructure and services that support sharing, access and re-use of open and linked data of agricultural research. It will allow research institutes in the area of agricultural statistics as well as in other areas of agricultural research open up their repositories of content and data and interlink and share them more effectively. The example related to agricultural statistics presented in this paper is only one of the numerous applications of the agINFRA products.

To achieve the aforementioned goals, current practices need to be overcome that produce information silos which lack accessibility and interoperability of the data resources. agINFRA promotes following Linked Data principles in order to remove such barriers. Furthermore the project devotes particular attention to the semantics of shared data as well as criteria of reliability such as data provenance.

Acknowledgement

The research leading to these results has received funding from the European Union Seventh Framework Programme (*FP7/2007-2013*) under *grant agreement* n° *283770*. We also thank all formal project partners and supporting organizations that combine their technologies, services and data in order to promote the open sharing of agricultural research results.

Corresponding author:
Dr Vassilis Protonotarios
Agro-Know Technologies
Grammou 17, 15235 Vrilissia, Athens, Greece
Tel.: +30 210 6897905, vprot@agroknow.gr

References

Androulakis, S. Buckle, A., Atkinson, I., Groenewegen, D., Nicholas, N., Treloar, A. and Beitz, A. *ARCHER – e-research tools for research data management.* International Journal of Digital Curation 4.1, 22-33, 2009 (on-line, accessed 23 November 2012), http://www.ijdc.net/index.php/ijdc/article/view/99/74

Aditya, T. and Kraak, M-J. *Aim4GDI: Facilitating the Synthesis of GDI Resources through Mapping and Superimpositions of Metadata Summaries.* Geoinformatica, Dec 2007, Vol. 11, Issue 4, pp. 459-478. DOI: 10.1007/s10707-007-0021-4, ISSN: 1573-7624, Springer, 2007

Benckiser G. and Schnell S. (eds) *Biodiversity in Agricultural Production Systems*. CRC Press, 2006.

Berners-Lee, T. *Linked Data* (on-line, accessed 17 November 2012), http://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., Heath, T. and Berners-Lee *Linked data - the story so far*. International Journal on Semantic Web and Information Systems (IJSWIS) 5.3, pp. 1-22, 2009.

Blake, M. *Economists Online: user requirements*. In: Serials 22(3), November 2009, http://eprints.lse.ac.uk/29914/1/Economists_Online_user_requirements_%28publisher%29.pdf

Chinese Crop Germplasm Research Information System (CGRIS), http://icgr.caas.net.cn/cgris_english.html

Clark, W., Tomich, T., van Noordwijk, M., Guston, D., Catacutan, D., Dickson, M. and McNie E. *Boundary work for sustainable development: Natural resource management at the Consultative Group on International Agricultural Research (CGIAR).* Proceedings of the National Academy of Sciences of the United States of America. Published online before print August 15, 2011, doi: 10.1073/pnas.0900231108, August 15, 2011 200900231 2011.

Coherence in Information for Agricultural Research for Development (CIARD), http://www.ciard.net

Descher, M., Feilhauer, T., Ludescher, T., Masser, P., Wenzel, B., Brezany, P., Elsayed, I., Woehrer, A., Tjoa, A.M. and Huemer, D. *Position paper: Secure infrastructure for scientific data life cycle management.* ARES 2009 Conference on Availability, Reliability and Security, paper available through the IEEE X digital library.

Dimitropoulos, A., Koutoumanos, A., Stoitsis, G., Sanchez-Alonso, S., Kastrantas, K. and Sicilia, M-A. *A community oriented approach providing truly multilingual access to agricultural learning objects*. Paper presented at the European Federation for Information Technology in Agriculture, Food and the Environment World Congress on Computers in Agriculture (EFITA 2011), 2011.

Economists Online (EO) portal, http://www.nereus4economics.info/projectneeo.html

European High-level Expert Group on Scientific Data *Riding the wave. How Europe can gain from the rising tide of scientific data*. A submission to the European Commission, October 2010 (on-line, accessed 17 November 2012), http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

European Roadmap for Research Infrastructures (ESFRI), *Report 2006*, Luxembourg (on-line, accessed 19 November 2012), ftp://ftp.cordis.europa.eu/pub/esfri/docs/esfri-roadmap-report-26092006_en.pdf

FAO AGRIS - International Information System for the Agricultural Sciences and Technology, http://agris.fao.org

FAO Fisheries and Aquaculture Department *ASFIS species classification* (on-line, accessed 19 November 2012), http://www.fao.org/fishery/collection/asfis/en

FAO Fishery Resources Monitoring System (FIRMS), http://firms.fao.org

FAO GeoNetwork, http://www.fao.org/geonetwork/

FAOSTAT, http://faostat.fao.org

Fogarolli, A., Brickley, D., Anibaldi, S. and Keizer, J. *AGRIS - From a Bibliographical Database to a Web Data Service on Agricultural Research Information.* Agricultural Information Management Worldwide, Vol. 4, No. 1, 2011. (on-line, accessed 23 November 2012), http://journals.sfu.ca/iaald/index.php/aginfo/article/view/196

Goff S. A., Vaughn M., Mckay S. *et al*. *The iPlant collaborative: cyberinfrastructure for plant biology.* Frontiers in Plant Science 2: 34, 2011 (on-line, accessed 17 Nobember 2012), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3355756/

Google - DSPL: Dataset Publishing Language (on-line, accessed 19 November 2012), https://developers.google.com/public-data

Gottron, T., Hachenberg, C., Harth, A. and Zapilko, B. *Towards a semantic data library for the social sciences.* Proceedings of the 1st International Workshop on Semantic Digital Archives (SDA 2011), Berlin, Germany, September 29, 2011. CEUR workshop proceedings, Volume 801, 2011 (on-line, accessed 24 November 2012), http://ceur-ws.org/Vol-801/paper4.pdf

Hey, T. and Hey, J. *e-Science and Its Implications for the Library Community*. In: Library Hi Tech, vol. 24, no. 4, 2006, pp. 515–28

Jarvis D.I, Padoch C. and Cooper D. (eds) *Managing Biodiversity in Agricultural Ecosystems*. Columbia UP, 2007.

Karampiperis, P., Manouselis, N. and Konstantopoulos, S. *Using a POWDER Triple Store for boosting the real-time performance of global agricultural data infrastructures.* Procedia Computer Science, Vol. 9, pp. 1578-1587, ISSN 1877-0509, DOI 10.1016/j.procs.2012.04.173, 2012.

Latin American Federation of Learning Object Repositories (LA FLOR), http://laflor.laclo.org

Michener, W. K. and Jones, M. B. *Ecoinformatics: supporting ecology as a data-intensive science*. Trends in Ecology and Evolution, February 2012, Vol. 27, No. 2 (on-line, accessed 23 November 2012), http://www.ncbi.nlm.nih.gov/pubmed/22240191

National Science Foundation Cyberinfrastructure Panel *Cyberinfrastructure Vision for 21st Century Discovery*. National Science Foundation, 2007 (on-line, accessed 17 November 2012), http://www.nsf.gov/od/oci/CI_Vision_March07.pdf

Organic.Edunet, http://www.organic-edunet.eu

Pesce, V., Maru, A. and Keizer, J. *The CIARD RING, an Infrastructure for Interoperability of Agricultural Research Information Services.* Agricultural Information Worldwide, Vol. 4, No 1, 2011 (on-line, accessed 23 November 2012), http://journals.sfu.ca/iaald/index.php/aginfo/article/view/213/170

Secretariat of the Pacific Community (SPC), http://www.spc.int

Statistical Data and Metadata Exchange (SDMX), http://sdmx.org

Research Papers in Economics (RePEc), http://repec.org

The Econometric Society, http://www.econometricsociety.org

Thessen, A. E. and Patterson, D. J. *Data issues in the life sciences*. In: Smith V. and Penev L. (eds) *e-Infrastructures for data publishing in biodiversity science*. Zookeys 150 (2011): Special issue: e-Infrastructures for data publishing in biodiversity science, 15–51 (on-line, accessed 23 November 2012), http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3234430/

Wezel A., Bellon S., Doré T., Francis C., Vallod D., David C. *Agroecology as a science, a movement and a practice – a review*. Agronomy for Sustainable Development, Vol. 29, Nr 4, 2009, pp. 503-515, http://www.agroeco.org/socla/pdfs/wezel-agroecology.pdf

VOA3R: Virtual Open Access Agriculture & Aquaculture Repository, http://voa3r.cc.uah.es

World Bank open data catalogue, http://data.worldbank.org/data-catalog