Yabowork, A., Orth, A., Ballantyne, P. 2017.
Making CGIAR outputs open and accessible: the CGSpace collaboration.
Knowledge Management for Development Journal 13(2): 23-33
http://journal.km4dev.org/

Making CGIAR outputs open and accessible: the CGSpace collaboration

Abenet Yabowork, Alan Orth, Peter Ballantyne

Abstract

In recent years, CGIAR centres and research programs have moved towards open access as part of commitments to make CGIAR information products widely accessible. These efforts span a wide variety of activities including adoption of policies, awareness raising, using open licenses and establishing open access repositories for products as well as data. This article explains the origins, operation and uses of the CGSpace repository set up in 2009 by the International Livestock Research Institute with several partners. Starting from an "institutional" effort, it has evolved into a collaboration among dozens of programs and entities, pooling technical efforts and generating collective public goods for the wider agricultural world. This article covers the CGSpace and open access value proposition, technical developments and choices, content management and standards, use and update, metrics and reach, as well as lessons and promising practices for wider use.

Keywords: institutional repository; agriculture; knowledge management; DSpace; CGIAR

Introduction

CGIAR (not an acronym) is a global research partnership for a food-secure future consisting of fifteen independent research centres. As highlighted in the CGIAR strategy, finding new and creative solutions to barriers to success is vital to deliver on the mission of the CGIAR[i]. Internal benchmarking studies of several CGIAR centres were conducted in 2008, with the results suggesting that the full texts of high-quality science produced by CGIAR centres, although identifiable, were not widely accessible[ii]. Later, in an independent review of the CGIAR, a panel of experts further emphasized the need for making CGIAR research outputs accessible[iii]. These studies and evaluations lead to the development of a "Triple A" framework by the ICT-KM program, as it became evident that, for CGIAR research benefits to travel across borders, they had to be Available, Accessible and Applicable[iv].

At about the same time, as part of the conversation about making CGIAR information products widely accessible, the International Livestock Research Institute (ILRI) launched a digital repository to capture the results of its research[v]. The platform, which would eventually come to be known as CGSpace, was based on a software application called DSpace. ILRI was quickly joined by a handful of other CGIAR centres and research programs in using CGSpace to begin

addressing the problems raised by the benchmarks. In this article "CGSpace" refers to the DSpace software application as configured by ILRI and its partners.

In 2013, all fifteen CGIAR research centres unanimously endorsed an *Open Access and Data Management Policy* designed to make final CGIAR information products like publications, datasets, and audiovisual materials Open Access. Open Access is considered one of the practical applications of commitment to this policy[vi]. Today, the CGSpace platform has become a flourishing collaboration between a dozen CGIAR centres, research programs, and partners, and its active use demonstrates a commitment to open access, open standards, and open licenses as a solution to help maximize the impact of their research.

This article is a reflection on the CGSpace value proposition to the CGIAR open access landscape. It covers technical developments, content management and standards, as well as lessons and promising practices for wider use.

CGSpace as a tool for the post-library digital information age

The DSpace software project was created in 2002 in a collaboration between Hewlett-Packard and the Massachusetts Institute of Technology[vii]. Now it is a free, open-source software project guided by a non-profit foundation and developed by a core group of software engineers along with a community of volunteers around the world. In retrospect, although it might not have felt like it at the time, when DSpace was released it was essentially at the crossroads of two eras that had come to be defined by an increasingly digital and connected world. For better or worse, traditional attitudes towards storing institutional knowledge in libraries were beginning to change, but digital systems for managing that knowledge were limited or non-existent.

ILRI was at the same crossroads in 2009 when it chose DSpace as the basis for its digital repository platform[viii]. Usage of the ILRI library had been evolving and there was increasing pressure to make information products digitally and widely accessible. Soon after the debut of CGSpace at ILRI, other CGIAR centres, research programs, and partners began joining the collaboration, citing the need for a modern repository that could be fit for purpose in the digital information age. This was not the first instance of DSpace in the CGIAR; it is the first so-called 'multi-tenant' installation.

CGSpace is a digital repository of agricultural research outputs and results produced by partners in collaboration. The repository indexes journal articles, reports, conference papers, proceedings, presentations, posters, videos, audio, policy briefs and more from across the CGIAR centres, research programs, and partners, in process making an enormous amount of agricultural information freely accessible for users around the world. Users can visit individual communities, search across the whole site, and sign up for email alerts and news feeds on topics that interest

them. CGSpace archives research outputs for referencing, reporting, and posterity, simultaneously serving both open access and publishing objectives.

It is important to note that many of the partners use CGSpace as an institutional index, as well as a repository. It does not just contain scientific articles and chapters, it has, or links to, multimedia products, posters, presentations, conference papers, and both internal and project documents. It is an index of 'grey literature' as well as formal products. This ensures a complete metadata record, facilitates reporting, and serves open access commitments. Globally, about one third of the content is not open access, some is just available for internal use with access restricted to institute staff), some are books and articles with copyright that do not allow access, and others are older legacy materials without digital versions. Since 2010, 75% or more of the content is open access.

CGSpace content management and standards

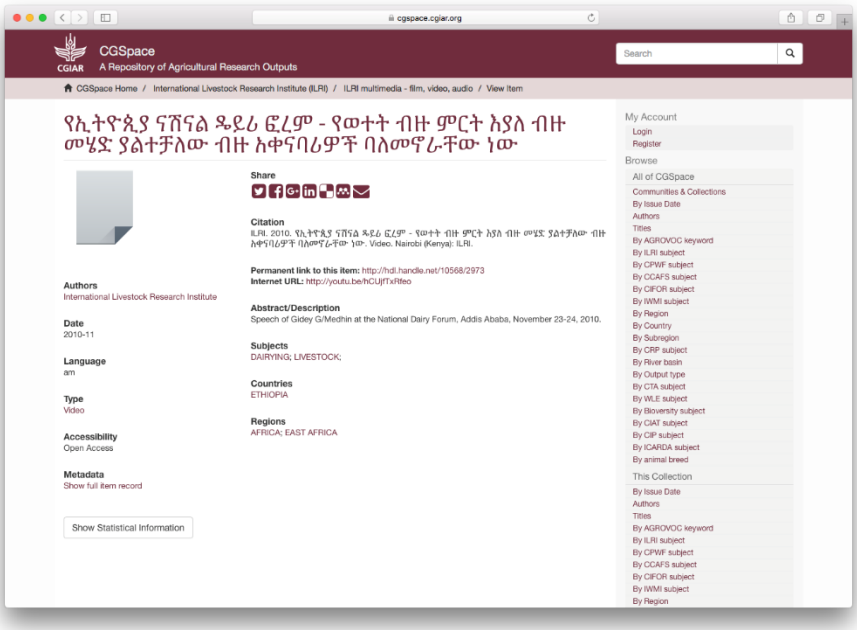
CGSpace organizes its content in three levels of hierarchy: communities, sub-communities, and collections (See: https://cgspace.cgiar.org/). At the very top are communities, which map more or less directly to each of the dozen or so CGIAR centres, research programs, and partners involved in the repository collaboration. Communities may contain sub-communities and collections, but content can only live at the bottom of the hierarchy inside collections. Content that logically belongs in more than one collection can be 'mapped' to other collections without having to make a duplicate accession. This functionality is used often on CGSpace, as it is common for CGIAR centres, research programs, and partners to collaborate on research, and it would be inefficient and incorrect for each of them to upload the same content.

The system of hierarchies is simple but powerful. In addition to their organizational capacity, communities, sub-communities, and collections also serve a practical administrative role. CGSpace allows assigning a range of permissions to users and groups within these hierarchical units. Each of these communities has their own administrators, for example knowledge managers or librarians at the respective organizations, and it is up to their discretion how to assign content submission, edit, and approval permissions to their users. The most common content submission workflow is where users fill out basic information in a submission form that gets reviewed, approved, or rejected by editors. Community administrators have these permissions and more, for example the ability to move, map, and delete content in their collections. This workflow adds several levels of quality assurance before content is published in the repository.

Metadata schemas for repository content, however, are managed globally. By default, the DSpace software project provides a rather barebones Dublin Core (DC) schema, but CGSpace has added a 'CG' schema developed through the wider CGIAR open data community to accommodate extra metadata fields where appropriate. For example, Dublin Core provides a

subject field (*dc.subject*) that is too general to put centre- and project-specific subject terms in, so CGSpace has adopted a convention of using qualified fields in the 'CG' schema like *cg.subject.ilri* and *cg.contributor.affiliation*. Technically, metadata itself is stored using Unicode (UTF-8) and therefore has no practical restrictions in supporting languages that use non-Latin character sets like Russian, Arabic, or Amharic (see image).

Figure: CGSpace item with Amharic language title and citation metadata (http://hdl.handle.net/10568/2973).



Data quality is a persistent issue that requires both user education and technical changes to submission workflows. Where possible, the use of "controlled vocabularies" has been a great help. For fields with a limited, official set of terms that don't change very often like country names, sponsors, and contributor affiliations, users are given a list of predefined values to choose from during content submission. Nevertheless, looking at repository data in aggregate can be extremely helpful, and batch data cleanup is necessary to normalize and correct errors in values for important fields. Workflows for data cleanup vary, but some tools for this include Excel, OpenRefine, and raw structured query language (SQL) commands in CGSpace's database[ix].

CGSpace technical architecture and strategy

The DSpace software suite is written in the Java programming language and makes use of either the PostgreSQL or Oracle database management systems. It is technically possible to run DSpace on any operating system environment that satisfies these requirements. When CGSpace was initially launched it was originally deployed on a Windows environment, but was quickly moved to a Linux system for security, speed, and to be more in line with web server management best

practices. Linux-based environments are vastly more suited towards scripting, automation, and batch data processing workloads.

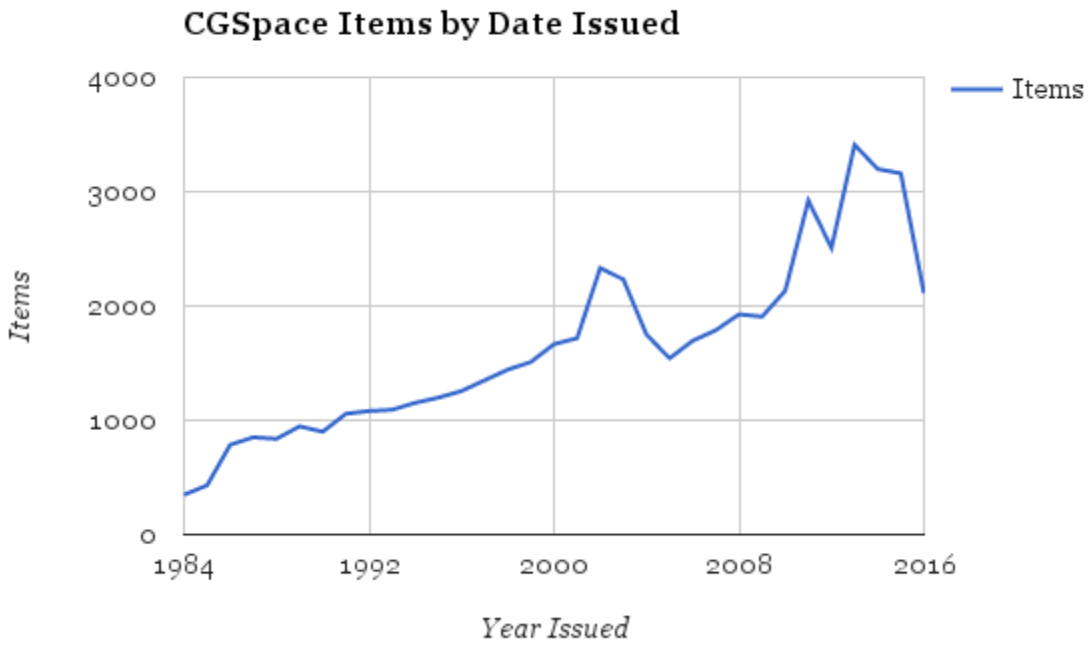
The DSpace community manages the development lifecycle of the application using a distributed source code management system called 'Git' via a web-based platform called GitHub. This platform hosts the code itself as well as provides a free, public place for developers to share code and log issues. CGSpace follows this convention and maintains a public, GitHub-based workflow for all customization of the platform[x]. The scope of customization is kept narrow to avoid drifting too far from the 'upstream' DSpace project and causing problems with integrating future features, improvements, and bug and security fixes.

CGSpace as a tool for durable preservation of digital content

In addition to managing information *about* institutional research outputs, CGSpace can manage digital versions of the outputs *themselves*. Digital content submitted to the repository is stored verbatim, without any modifications or proprietary data storage formats locking it into the platform. Should the need arise to move or migrate repository contents to another system, CGSpace provides several open, industry standard interfaces to programmatically query and export data.

Content is typically submitted individually as it becomes available, but it is also possible to submit content in batch. One common use for batch uploads—and a major success story—is the digitizing and uploading of old, physically archived information products from CGSpace partner institutions. Despite having only been launched in 2009, CGSpace currently contains over 17,000 of such items from the 1980s, 1990s, and early 2000s. Some of this content is metadata only, many link to full digital content on Google Books or elsewhere, but many are accompanied by digitized copies from institutional archives. Thus, in 2016, ILRI completed the migration of all its archival materials, including from its predecessors the International Livestock Centre for Africa and the International Laboratory for Research on Animal Diseases.

Figure: CGSpace items by date issued



Repository content is backed up regularly to an off-site location.

CGSpace as a tool for publishing and dissemination

One of the features that sets modern digital repositories like CGSpace apart from traditional library management systems is their ability to be queried programmatically. CGSpace supports a number of industry standard query interfaces like REST and OAI-PMH. Search engines use these interfaces to crawl, consume, and index repository content in a structured, machine-readable way. Software developers from Google have gone even further, working with the DSpace software project to add meaningful metadata markup that helps Google Scholar make more sense of repository content.

The same interfaces that search engines use to crawl the repository can be used to disseminate content in other ways too. Several partners in CGSpace are harvesting and displaying their content across other platforms. The CGIAR Research Program on Climate Change, Agriculture and Food Security (CCAFS), for example, has developed a website showcasing their publications, including all metadata, high-quality thumbnails, and a link to the original item accession on CGSpace[xi]. Similar approaches, with different technical solutions, have been adopted by the CGIAR Research Program on Water, Land and Ecosystems[xii] and the

International Center for Tropical Agriculture[xiii]. CGSpace content about Ethiopia is being harvested and made accessible through the Ethiopian Agricultural Portal[xiv] and there are plans for the whole site to be harvested through AGRIS[xv]. Less technically sophisticated, CGSpace also supports simple syndication protocols like RSS[xvi] and Atom that enable content feeds to populate blog sidebars, social media, and user email inboxes.

These kinds of integrations are important because they avoid creating duplicate versions of content across many sites, and instead defer to CGSpace as the sole, authoritative source for institutional information products.

CGSpace as a tool for collaboration and value addition

As indicated earlier, CGSpace was first set up by ILRI to support its institutional needs. Very early on, it was apparent that much of the content held by ILRI was not totally of ILRI. Thus the CGIAR System-wide Livestock Programme (SLP) was hosted by ILRI and produced research outputs from six or seven other collaborators that could not in all honesty be attributed to ILRI. Thus, the first need for a platform to host and make visible not just ILRI content but also SLP content emerged. Initial discussions with Bram Luyten working at Atmire[xvii], a DSpace specialist company, suggested that a 'multi-tenant' approach to DSpace was feasible. By this, the aim was to present content from different entities as communities on one interface, also allowing content in a specific community to be curated and managed independently.

Once this approach was in place, other partners seeking an open-access repository solution joined the collaboration[xviii]. The initial drivers to collaborate were by initiatives like the CGIAR Challenge Program on Water and Food (CPWF) and the newly established CGIAR research programs (CCAFS, WLE, Livestock and Fish) that needed platforms that could store and display and publish information products beyond institutional boundaries. There was a strong interest in sharing costs and technical infrastructure and to collectively build a resources that is greater than the sum of the parts. This process was not without hiccups: some partners joined but then decided they wanted to run their own platforms. Others kept their legacy library catalogues and joined, seeing CGSpace as an additional vehicle to make their products visible and accessible. Still others joined but, for various reasons, never invested in the platform as their primary repository.

In summary, the motivations and depths of engagement vary across the partners. The collaboration has, however, allowed this to happen with different partners joining at their own speeds and as their own capacities allow.

What were the primary benefits of this collaboration? For the various partners, it has allowed them to have a robust and affordable repository to publish and archive their products. For many,

the cost-sharing elements have sustained a robust open source technical platform and support at levels that would not be feasible in a single organization, and indeed would not be desirable to replicate multiple times. The participants have also been able to build on collective learning and tips, drawing in expert technical inputs from Atmire, for example, co-financing additional modules, subscribing collectively to Altmetric, and taking advantage of each DSpace upgrade in a consistent manner. The collective size and frequent updating of the repository also ensures it is well-indexed in global search engines.

A less tangible benefit, though very interesting, is the potential the shared platform offers to see patterns and trends in the collective content. The site is perhaps the only public place where views across institutes and programs are visible. Frequencies of authorship (and co-authorship), author affiliations, CGIAR research program associations, investor frequencies, information product type distribution, and geographic focus are all visible across the site as a whole or specific collections, revealing interesting insights into CGIAR research publishing trends. This is incomplete so not as powerful as it could be, and some visualization modules in development will hopefully show the patterns in more powerful ways.

CGSpace metrics

Gaining insights into use and uptake is an increasingly important aspect of knowledge management. Besides detailed usage and user tracking, researchers are interested in the use of 'their' products; research managers want to track and report usage for projects or priority countries, and CGIAR centres and research programs are more and more being asked to show the reach and even impact of their work.

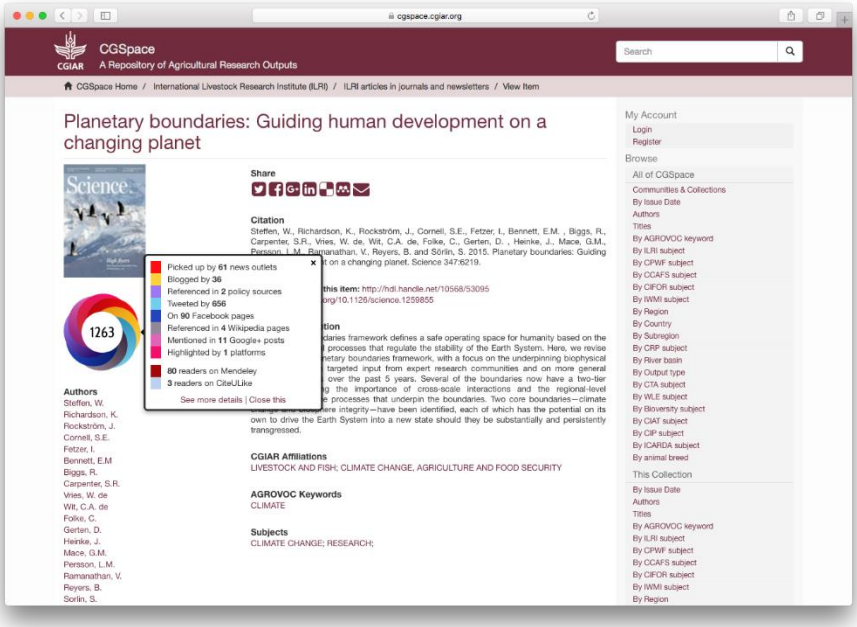
As in other DSpace institutional repositories, CGSpace offers standard statistics which includes views and downloads for communities, sub-communities, collections and items. The standard statistics also give further information about most popular authors and items in the entire repository or in specific collections; it also gives statistics on views and downloads by country. In addition, CGSpace has add-on modules to get more detailed reports on content. These include the Atmire-provided content and usage analysis; listings and reports; and metadata quality modules which are a source of important information for managers. On top of these, Google Analytics and Altmetric give additional and alternative metrics on items published in the repository. Although it is still not easy to quantify the impact of the research made accessible through CGSpace, there is an evidence of measurable increase on views and downloads of items hosted on the CGSpace.

Table of views and downloads

Community	2011	2012	2013	2014	2015	2016	All time
Africa RISING	9	61,411	96,752	107,399	153,365	237,335	656,270
Bioversity International	93	289	513	9,829	44,139	180,946	235,809
CIFOR		189,392	248,901	211,026	51,063	51,632	752,014
CPWF	145,436	411,801	486,026	672,592	450,484	373,338	2,539,677
CCAFS	16,277	76,157	160,130	260,300	477,317	780,882	1,771,063
L&F	13,426	45,939	71,436	142,540	257,493	420,413	951,246
WLE	368	6,732	14,283	103,173	156,432	203,660	484,647
SLP	116,152	107,987	80,180	84,004	44,108	99,677	532,108
CIAT	2,081	5,404	9,243	147,317	1,232,841	2,275,383	3,672,270
CGIAR Gender and Agriculture Research Network		4,746	7,194	8,628	58,430	92,293	171,291
IITA	69	2,404	4,788	5,788	11,290	68,525	92,864
ILRI	1,810,140	2,987,078	2,994,837	3,231,553	3,399,545	4,571,245	18,994,398
CIP	175	1,371	1,512	2,847	41,126	97,881	144,913
IWMI	76	417	642	115,574	69,303	42,948	228,961
CTA	1,057	1,126	851	48,571	165,363	327,715	544,682
Others	85,229	121,266	106,569	119,738	78,772	121,559	458,874
TOTAL	2,190,589	4,047,842	4,321,298	5,343,325	6,856,552	10,157,283	32,916,889
Change on previous year		185%	107%	124%	128%	148%	

Beyond the metrics and statistics provided by the site, in late 2015 some of the CGSpace partners decided to use Altmetric[xix] to give additional insights into social attention given to different items. Subscribing to Altmetric provides access to a dashboard where attention to specific items, authors and collections can be checked. For items that are getting social media attention (Twitter, blogs, Facebook, Wikipedia, etc.), reports and insights can be generated. Altmetric also tracks when items are cited in policy documents and Scopus, so it provides a powerful way to go beyond repository views and downloads and beyond journal 'impact' factors.

Figure: CGSpace item with Altmetric data showing social attention



Lessons and insights

CGSpace offers a robust way for participating partners to: publish and archive research and other outputs, record outputs for reports; re-publish outputs across other platforms and track metrics on hosted outputs. It meets open access commitments.

For ILRI, it has provided the pathway to publish all of its research outputs through one channel, getting them off websites and properly curated. This publishing role has proved to be much more useful that a repository role as a value proposition to scientists and managers in the institute.

CGSpace is a good way for information seekers and users to search, find and gain access to the research and other outputs of participating partners.

Sharing a technical platform is feasible and affordable (global common costs, usually with extra inputs from services such as Atmire and Altmetric cost approximately USD90,000 per year). Decentralizing content curation among partners is effective and practical. Sharing the technical load has allowed the partners to focus on content, curation and dissemination. Within institutions, information and communication specialists are the main content managers; having scientists submit their own content is less frequent—perhaps as the institutes want to 'manage' submissions and metadata more tightly.

The path to keep CGSpace technical development as close to the mainstream DSpace code and

development as possible allowed the project to easily upgrade and take advantage of new features. Other DSpace instances where the code has been modified often struggle to keep up.

There is a lot of work involved in setting up, standardizing and keeping the metadata and taxonomies consistent and relevant. The CGSpace approach has been to use global standards like Dublin Core fields and AGROVOC taxonomy, together with evolving CGIAR-wide 'CG-core' fields and a few locally-demanded field of partners for specific needs.

The user interface to DSpace is not always fully appreciated by users. This is being improved with each upgrade but some 'off-site' interfaces often provide better results. What is important is to make sure that the item 'handle' addresses are always retained as these act as permanent URLs and are the basis for services like Altmetric to track attention.

Generating reports, especially management ones, is a bit cumbersome and it is hoped that the DSpace project as a whole may give more attention to this.

Partnering with a platform specialist, in this case Atmire, was critical in the early days as it provided training, technical and related support. As the platform progressed, the collaboration was more critical on the technical side with Atmire providing inputs based on their close involvement in the global DSpace community.

The CGSpace collaboration was built from the bottom, working closely with people with shared goals and objectives. This proved effective in implementation and avoided any need to push, pull or coerce all CGIAR entities into using it.

About the authors

Abenet Yabowork is Knowledge Curation Manager at the International Livestock Research Institute (ILRI) in Ethiopia. In this role, she manages ILRI's corporate repository on CGSpace and supports other partners in using the platform. Before joining ILRI, she has worked as a website developer and public information assistant at the International Labour Organization (ILO) in Ethiopia. She has a Bachelor of Science in Information Systems from Addis Ababa University. Email: a.yabowork@cgiar.org

Alan Orth is a freelance consultant based in Bulgaria. He currently manages the technical workflows of the CGSpace repository. He was previously Linux systems administrator at the International Livestock Research Institute (ILRI) where he managed the research-computing platform and before that, he worked as a volunteer teaching computer science at a rural college in Kenya. He has a Bachelor of Science degree in Computer Information Systems from the California State University, Chico. Email: aorth@mjanja.ch

Peter Ballantyne is Head of Communications and Knowledge Management at the International Livestock Research Institute. Before ILRI, he was Director of 'Euforic' an information-sharing cooperative working to enhance access to information and knowledge on Europe's international development. He also worked as a freelance agricultural information specialist with the CGIAR ICT-KM Program, DFID, several CGIAR centers, FAO, GFAR and others. He started his career working with agricultural information and knowledge management – first at the World Bank, then at a Faculty of Agriculture in Thailand, then in the CGIAR at the International Service for National Agricultural Research (ISNAR). After ISNAR, he mainly worked in the international development sector, with ECDPM, IICD, and INASP. Email: p.ballantyne@cgiar.org

---

[i] http://hdl.handle.net/10947/4069

[ii] Arivananthan, M.; Ballantyne, P.; Porcari, E.M. 2010. Benchmarking CGIAR research outputs for availability and accessibility. *Agricultural Information Worldwide* 3(1): 17-22. http://hdl.handle.net/10568/1554

[iii] CGIAR Independent Review Panel. 2008. Bringing Together the Best of Science and the Best of Development. Independent Review of the CGIAR System. Report to the Executive Council. Washington, DC. http://hdl.handle.net/10947/4949

[iv] Ballantyne, P.G. 2008. Making CGIAR research outputs available and accessible as IPGs. http://ictkm.cgiar.org/document_library/program_docs/ICT-KM%20AAA_complete.pdf

[v] https://maarifa.ilri.org/2009/11/25/ilri-2-0-update-on-ilri-web-communications/

[vi] https://www.cgiar.org/open

[vii] http://www.dlib.org/dlib/january03/smith/01smith.html

[viii] Ballantyne, P.G. and Keizer, J. 2009. Evaluation of the ILRI InfoCentre: Report of a Center-Commissioned external review. Nairobi: ILRI. http://hdl.handle.net/10568/78392

[ix] https://github.com/ilri/DSpace/wiki/Scripts

[x] https://github.com/ilri/DSpace

[xi] https://ccafs.cgiar.org/publications

[xii] https://wle.cgiar.org/resources/publications

[xiii] https://ciat.cgiar.org/data-information-knowledge/ciat-research-online

[xiv] http://www.eap.gov.et/resources

[xv] http://agris.fao.org

[xvi] ILRI project websites and blogs all include feeds from CGSpace — see example at https://news.ilri.org

[xvii] https://atmire.com

[xviii] https://cgspace.cgiar.org/page/about has some of the background and history.

[xix] https://www.altmetric.com