CGIAR RESEARCH PROGRAM ON Water, Land and Ecosystems

LED BY: IWMI International Water Management Institute

LEGACY BRIEF 2

Sharing research data in a connected world for connected decisions: Lessons learned from the CGIAR Research Program on Water, Land and Ecosystems (WLE)

Trends in data science demonstrate the utility of big, open access data in addressing global challenges such as land degradation, food and nutrition security, biodiversity loss and climate change. As vastly more data becomes accessible, new skill sets are required in order to collect, manage, curate, store, share and analyze these datasets. This brief provides lessons learned on collecting, managing and utilizing research data from a 10-year research-for-development program that aimed to advance research and innovation in the areas of soil, land, water and ecosystems. While institutional changes will take time, this paper advocates an integrated approach as a more immediate solution for managing large amounts of research data – from the outset of projects until the end with a specific focus on data ethics and reusability. Resources, capacities and organizational structures and norms must be built around these objectives.

Management of research data at WLE

Management of research data is central to research-for-development organizations. Over the last 10 years (2012–2021), partners of the CGIAR Research Program (CRP) on Water, Land and Ecosystems have continuously adapted their internal processes for managing research data in response to emerging trends and in adherence to the CGIAR Open Access and Data Management Policy (CGIAR 2013), recently replaced by the CGIAR Open and FAIR Data Assets Policy (CGIAR 2021).

Research Officer working on water data in Kathmandu (*photo:* Sharad Maharjan / IWMI)

WLE partners have integrated data across geographies, themes, disciplines and data types – advancing systems working and measuring aspects in fields such as soil, land, water, biodiversity, agrobiodiversity, human wellbeing and governance using a plethora of collection methods (e.g. rapid surveys, geographic information systems, participatory mapping, field sensors, water gauges, focus group discussions). The result of this data management is illustrated in Figure 1, which shows some WLE publications according to year and theme throughout Phase II of the program. However, the existing systems for managing research data must be further improved to ensure that data is truly 'FAIR' (i.e. findable, accessible, interoperable, reusable) and can be widely used to inform timely policy- and decision-making processes in the future (Deutz et al. 2020).

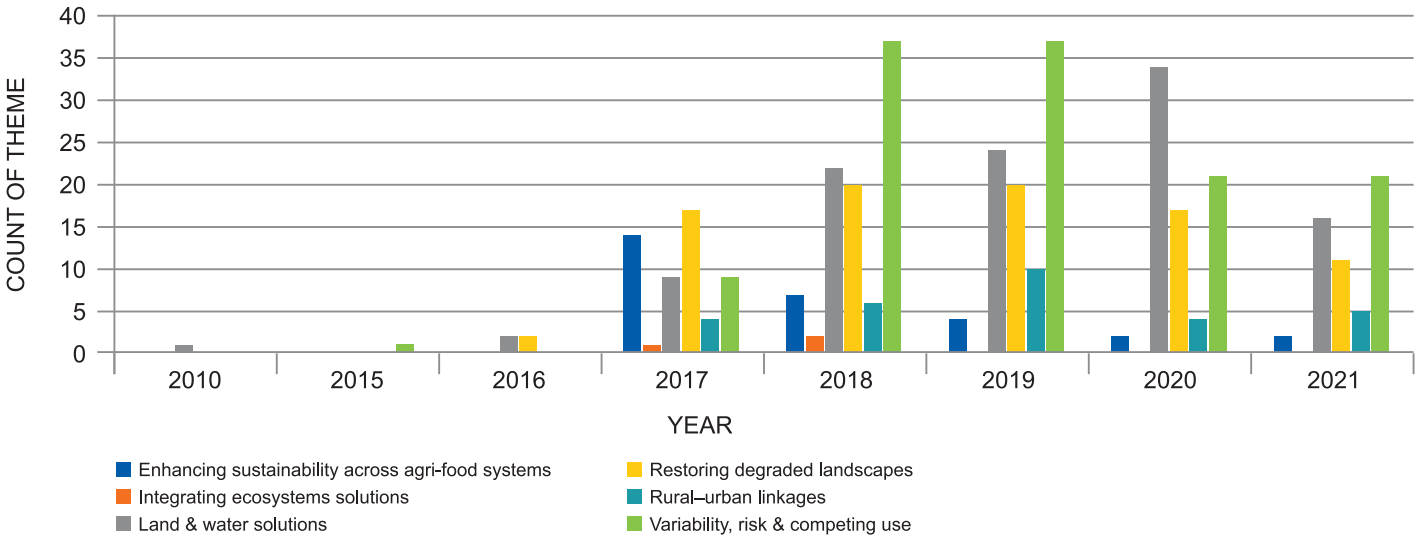

Figure 1: WLE publications per year and theme throughout Phase II of the program, based on data from CGSpace (note that not all centers upload publications to CGSpace so the data shown is incomplete).

This brief provides lessons learned on collecting, managing and utilizing research data from WLE's 10-year research-for-development program that aimed to advance research and innovation in the areas of soil, land, water and ecosystems. The main objectives of this paper are to: 1) Briefly outline key opportunities and potential for open access data on soil, land, water and ecosystems; 2) Communicate experiences on sharing data across multiple research institutions and disciplines; and 3) Provide recommendations on the way forward.

Trends in data science: Managing open access and interdisciplinary datasets

Trends in applied research indicate that the utility of big, open access datasets to answer key critical questions is on the rise, as researchers gather evidence on topics such as land degradation, food and nutrition security, biodiversity loss and climate change. The emerging opportunities are accompanied by advances in electronic hardware and software as well as thriving fields such as artificial intelligence, machine learning and natural language processing.

These developments, among others, bring challenges for the management of research data in any organization. Global analyses require compatibility across datasets and there is a growing interest in transdisciplinary research. This includes data analytics in the agricultural sector, where opportunities to combine multiple data sources can benefit the stakeholders involved (Kamilaris et al. 2017). A review of big data in smart farming demonstrated that the big data revolution could transform interdisciplinary agricultural research by combining multiple aspects of the agricultural value chain (Wolfert et al. 2017). Open access to curated, high-quality data is at the core of the big data revolution. In fact, sectors that commit to generating and sharing their datasets are more likely to be at the forefront of groundbreaking scientific developments.

As vastly more data from multiple sources becomes accessible, new skill sets are required in order to collect, manage, curate, store, share and analyze these datasets. Merely generating data with the hope that others will reuse it is not enough to utilize the existing potential. Investment in the long-term sustainability of datasets is necessary (Bourne et al. 2015). In addition, attention must be paid to the design and functionality of the data ecosystems through which we channel our knowledge and address questions on complex ecological systems (Welle Donker and van Loenen 2017; Wolfert et al. 2017).

Data ecosystems consist of actors, roles, relationships and resources in a shared network. Rather than incorporating a common platform, they build on a wide collection of data-based resources and are self-regulating – as competition and collaboration regulate actors and resources through the common interest of creating value from data (Oliveira and Lóscio 2017). Data ecosystems therefore shape the way data-based knowledge is collected, managed, analyzed and shared and are influenced by relevant norms on these activities and vice versa. For this reason, it is essential to adapt practices around research data to match developments in data science and the global challenges ahead. Institutional changes in this area are already under way but they will take time and must build on, and be guided by, best practices such as the FAIR principles.

Lessons learned and recommendations for effective data management

WLE researchers and managers reflected on obstacles to, and best practices for effective management of research data as experienced over the last decade. The program has learned from these reflections and shares arising recommendations below, with the objective of informing future decisions on managing research data – decisions made by CGIAR as well as other integrated research-for-development programs.

Coordinate and incentivize an integrated approach to managing research data within and across projects from start to finish, and provide a foundation for ethical reutilization in the future. Too often, research data has only been harvested when a member of staff leaves a research organization, at the end of projects, or when required by a journal. A consistent approach to data harvesting from the outset of most WLE partner projects has been lacking; this has resulted in WLE partners either being left with some non-curated datasets not ready for upload, or losing research data completely. Generating relevant metadata under these conditions is difficult. In recent years, some CGIAR Research Centers have established data curation units to tackle this issue (among others) but processes still need adjustment. Project planning and management processes need to incorporate data management throughout the project cycle, using a series of practical steps, as suggested below:

Stage 1: Project strategy/proposal and launch

- Define and implement data ethics that reflect the positionality of all stakeholders and thereby respect, empower and capture multiple voices. Base such data ethics on information from marginalized groups and key stakeholders (e.g. farmers) to avoid further supporting power imbalances and injustices.

- Develop a research data management plan from the outset covering:
  - Types of data, as well as metadata, to be collected – focusing on reusability (e.g. embed data in global assessments, replicate data collection in other locations) and entailing ontologies (CGIAR Platform for Big Data in Agriculture n.d.) for harmonizing measurements, metrics and information.
  - How data will be managed including necessary approval processes (institutional review boards or similar) and data and standards, e.g. International Organization for Standardization (ISO), European Union, FAIR, ROSES – the RepOrting standards for Systematic Evidence Syntheses (ROSES 2017).
  - Where, when, by whom and according to what guidelines data will be uploaded.
  - How datasets and findings will be shared internally and externally, and how often.

- Make sure to allocate sufficient funding to ensure not only that data and metadata is FAIR but also that the results produced are understandable to all stakeholders and rights holders involved.

- Include good data practices in project/staff performance metrics.

Stage 2: Project implementation

- Invest in human and technology capacity development across the sites and contexts where data is being harvested, so that collected data can effectively guide local decisions whilst informing global efforts.

- Make sure that large and/or data-intensive projects regularly upload data to the relevant storage system/platform and update metadata accordingly. To ensure compliance, project managers could use financial incentives such as releasing funds upon successful upload.

- Have knowledgeable curators to ensure that datasets are high-quality and well-curated to increase reusability after uploading. This is especially important for qualitative data typically requiring more curation.

Stage 3: Project closure

- Incentivize uploading of any remaining unpublicized datasets; for example, by setting FAIR research data as a requirement to close a project with a process in place to send regular reminders until a project is closed.

- Finalize metadata and share it with relevant entities (e.g. program management, partners).

- Share the collected data as well as the analyses that it supported (ideally) with every person, institution and organization involved in the data production process in an understandable format.

- Evaluate compliance with data practices and ethics set out in the research data management plan.

Implement data ethics that promote equitable representation and reflect power imbalances and injustices. Organizations and researchers have too often perceived data collection as a routine process in which participants' knowledge is valued but opinions on the data itself are rarely considered. Hence, insufficient time and effort is put into engaging with data or knowledge providers (e.g. indigenous communities, farmers, women, youth) to consider how shared data might make them more vulnerable and to better understand and discuss what data is really needed in local, regional and global contexts. Reflecting on positionalities and power imbalances, researchers should see data and knowledge providers through an intersectional lens that captures human diversity and represents all voices equitably. This not only makes research more ethical but also improves its quality and relevance. More inclusive planning around data collection could also support an organization's gender and inclusion goals. Strengthening inclusion and equality in data collection and management should build on the processes for ethical approval that many institutions have put in place already; efforts are needed to integrate ethics and inclusion aspects more strongly in the future, as argued in Data Feminist (D'Ignazio and Klein 2020).

Ensure effective responsibility over research data and information sharing among entities in the same project, and harmonize attribution and upload processes. WLE partner centers have been responsible for the manipulation, harmonization, uploading and storing of WLE data. These CGIAR Research Centers each have in place their own data systems, some of which are advanced while others are still being built and undergoing regular changes. Therefore, it has been difficult for the management staff of the CRP to know and control processes of research data management, including attribution to WLE in published datasets and articles. This has contributed to several WLE publications not being published on CGSpace, and limited oversight of the CRP's uploaded research data. It follows that effective communication pathways for sharing project information across the included entities are crucial for effective management of research data (e.g. partners informing CRPs when uploading data). Attributions and upload processes therefore need to be harmonized through guidance on data attributions and information sharing for authors (e.g. including all funding entities in the acknowledgement section of articles). To prevent scattering of research data and publications, widely used platforms such as Harvard Dataverse, figshare or Mendeley Data are recommended, although institutional repositories have the advantage of being tailored to a specific organization.

Allocate sufficient resources and build capacities for the collection, management, curation, analysis and communication of research data and related findings, both locally and institutionally, and boost reusability. In the past, resources and capacities have often been insufficient for managing and publishing research data (and metadata) appropriately. Budgets and internal capacities must enable projects to ensure that all research data – regardless of scale, form or topic – are FAIR. In particular, reusability should be a key objective in future research. The following practical steps are suggested:

- Account for the management, collection, curation and analysis of research data when making decisions on resources and staffing, acknowledging how much time is needed for these and related tasks. This may require increased online security measures and storage. Project teams could include a person (possibly full-time) who is continuously working on the management of research data, including its FAIR upload and sharing. Adhering to the FAIR principles should be included in the staff's terms of reference.

- Allocate funding and time for sharing data with, and communicating research findings to, data and information providers in a way that is easy to understand (e.g. through creative data visualizations). In doing so, researchers might consider alternative means of presentation such as art, storytelling or games, as appropriate in the context.

- Make sure that projects/researchers prioritize data curation sufficiently. While poor data quality can sometimes be an issue causing researchers to not upload their data, it is proper data curation that is more often lacking. Good data curation enables data sharing and reuse so that other scientists and researchers can reproduce results and perform further experiments based on the same data. In particular, the sharing of qualitative data, which is often not considered to be publishable, can be enhanced through data curation.

- Build capacities through skills training for researchers where needed (e.g. in data science, data ethics, artificial intelligence, machine learning, natural language processing and transdisciplinary research) and/or expect relevant skills when hiring. Staff responsible only for managing research data rather than collecting it still need the skills to assess data quality and its adherence to standards.

- Boost the reusability of research data by embedding it in global assessments from the planning stage, and build on existing templates for data collection, possibly replicating existing cases. Harmonizing uploaded

and/or published datasets gives further opportunities for hypothesis testing as well as other types of cross-site and cross-divisional analysis. Above all, making data reusable respects the time and effort of respondents, communities and other stakeholders involved in the data collection process and promotes accountability and learning among researchers and implementing organizations.

Increase the visibility of research data and promote a mindset among researchers and partners that incentivizes effective implementation of the FAIR principles. In the past, researchers sometimes had a mindset of 'treasuring up' their research data: a reluctance to share data before the end of a project or the publishing of a related article. If not urged to do so, many did not publish their data at all. The prospect of leaving an organization can add to a lack of motivation to share and upload data. Research data that is already published is, in turn, often not sufficiently promoted inside and outside of an organization, further disincentivizing publication and hindering reuse. In response to this, successfully published research data should be acknowledged and showcased internally (e.g. via emails, meetings, prizes for the most published datasets, blogs) and play a role in the performance evaluations of staff. More generally, communications can highlight open access databases and related publications, such as Our World in Data (Global Change Data Lab n.d.). Data publications can also be promoted externally on a relevant website with a picture of the researchers involved. Both data articles and datasets should count as a publication (with a DOI for citations) and be included in the publication list of the center/program/project supporting the (re-)use and analysis of said data.

Conclusions

Increasingly, both policy and investments are based on evidence. Underpinning evidence are data. Therefore, the sound management and curation of data, as advocated in this brief, are fundamental entry points to sound policy and investment decisions.

It is clear that effective management of research data requires more than additional funding. Structural and cultural changes in the way research projects are designed and conducted are necessary. These changes require training and mentorship and it will take time for them to manifest in institutions and their staff. In this process, it should be recognized that the five lessons learned above are interrelated, and that synergies between them can accelerate change. Ideally, processes are harmonized across projects and organizations, and the benefits of FAIR and locally/globally relevant data are experienced by researchers, partner organizations and other stakeholders, further propelling the transition.

This reflection piece comes out of WLE researchers' experiences over 10 years. We hope that the lessons learned will inform and enable forthcoming research initiatives of One CGIAR or other large-scale integrated research-for-development programs to transition towards managing data in a way that informs and empowers local communities, countries and other stakeholders (e.g. private sector, non-governmental organizations) to make evidence-based decisions. The recommendations illustrated above are a first step in this direction.

Researchers collect water level data from a data logger (*photo:* Faseeh Shams / IWMI).

References

Bourne, P.; Lorsch, J.; Green, E. 2015. Perspective: Sustaining the big-data ecosystem. *Nature* 527: 16-17. DOI: https://doi.org/10.1038/527S16a.

CGIAR. 2013. *CGIAR Open Access and Open Data Policy*. Available at: https://www.cgiar.org/how-we-work/accountability/open-access/.

CGIAR. 2021. *CGIAR Open and FAIR Data Assets Policy*. Available at: https://cgspace.cgiar.org/bitstream/handle/10568/113623/CGIAR_OFDA_Policy_Approved_16April2021.pdf?sequence=1&isAllowed=y.

CGIAR Platform for Big Data in Agriculture. n.d. *Ontologies. Community of Practice*. Available at: https://bigdata.cgiar.org/communities-of-practice/ontologies/.

Deutz, D.B.; Buss, M.C.H.; Hansen, J.S.; Hansen, K.K.; Kjelmann, K.G.; Larsen, A.V.; Vlachos, E.; Holmstrand, K.F. 2020. *How to FAIR: a Danish website to guide researchers on making research data more FAIR*. Available at: https://doi.org/10.5281/zenodo.3712065.

D'Ignazio, C.; Klein, L.F. 2020. *Data feminism*. Cambridge, MA, USA: The MIT Press. Available at: https://data-feminism.mitpress.mit.edu/.

Global Change Data Lab. n.d. *Our World in Data*. Available at: https://ourworldindata.org/.

Kamilaris, A.; Kartakoullis, A.; Prenafeta-Boldú, F.X. 2017. A review on the practice of big data analysis in agriculture. *Computers and Electronics in Agriculture* 143: 23-37. DOI: https://doi.org/10.1016/j.compag.2017.09.037.

Oliveira, M.I.S.; Lóscio, B.F. 2018. What is a data ecosystem? In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age (dg.o '18)*. New York, NY, USA: Association for Computing Machinery. Article 74, pp. 1-9. DOI: https://doi.org/10.1145/3209281.3209335.

ROSES n.d. *About ROSES*. Available at: https://www.roses-reporting.com/about-roses.

Welle Donker, F.; van Loenen, B. 2017. How to assess the success of the open data ecosystem? *International Journal of Digital Earth* 10: 284-306. DOI: https://doi.org/10.1080/17538947.2016.1224938.

Wolfert, S.; Ge, L.; Verdouw, C.; Bogaardt, M. 2017. Big data in smart farming – a review. *Agricultural Systems* 153: 69-80. DOI: https://doi.org/10.1016/j.agsy.2017.01.023.

Acknowledgments

This research was carried out as part of the CGIAR Research Program on Water, Land and Ecosystems (WLE) and supported by Funders contributing to the CGIAR Trust Fund (www.cgiar.org/funders/). CGIAR is a global research partnership for a food-secure future.

Contacts

Jonathan Wirths, International Water Management Institute (IWMI); Leigh Winowiecki, World Agroforestry (ICRAF); Natalia Estrada-Carmona, Alliance of Bioversity International and the International Center for Tropical Agriculture (CIAT); Emma Greatrix, IWMI; Simon Langan, IWMI; Stefan Uhlenbrook, IWMI (wle@cgiar.org).

Suggested citation

Wirths, J.; Winowiecki, L.; Natalia Estrada-Carmona, N.; Greatrix, E.; Langan, S.; Uhlenbrook, S. 2021. *Sharing research data in a connected world for connected decisions: Lessons learned from the CGIAR Research Program on Water, Land and Ecosystems (WLE)*. Colombo, Sri Lanka: International Water Management Institute (IWMI). CGIAR Research Program on Water, Land and Ecosystems (WLE). 6p. (WLE Legacy Brief Series 2)

CGIAR RESEARCH PROGRAM ON Water, Land and Ecosystems

LED BY: IWMI International Water Management Institute

The CGIAR Research Program on Water, Land and Ecosystems (WLE) is a global research-for-development program connecting partners to deliver sustainable agriculture solutions that enhance our natural resources – and the lives of people that rely on them. WLE brings together 11 CGIAR centers, the Food and Agriculture Organization of the United Nations (FAO), the RUAF Global Partnership, and national, regional and international partners to deliver solutions that change agriculture from a driver of environmental degradation to part of the solution. WLE is led by the International Water Management Institute (IWMI) and partners as part of CGIAR, a global research partnership for a food-secure future.

CGIAR Research Program on Water, Land and Ecosystems

International Water Management Institute (IWMI)
127 Sunil Mawatha, Pelawatta
Battaramulla, Sri Lanka
Email: wle@cgiar.org
Website: wle.cgiar.org
Thrive blog: https://wle.cgiar.org/thrive

IN PARTNERSHIP WITH: Bioversity International, CIAT, FAO, ICARDA, ICRISAT, IFPRI, RUAF, World Agroforestry, CIFOR, CIP, ILRI, WorldFish