GODAN

Global Open Data
for Agriculture & Nutrition

A Global Data Ecosystem for Agriculture and Food

FOREWORD

There is widespread and justified concern around the immense challenges to food systems in meeting the demand of future populations for sufficient, affordable, safe and nutritious food. There are equal concerns about meeting those challenges in ways that both reduce unintended negative environmental consequences and make significant contributions to the livelihoods of farm families and promote inclusive national growth. Making effective progress in such a complex and dynamic setting, in an era where threats and opportunities appear to grow in equal and rapid measure, demands urgent, coordinated, and transformative action in better marshalling and exploiting revamped data infrastructure across the food system landscape.

This report is both a timely and a critical exploration of a daring vision of a global data ecosystem for agriculture and food that no single entity or network could possibly seek to deliver, but that might guided by principles presented by the authors - lie within the grasp of an open and adaptive global stakeholder coalition. Daring visions are inevitably replete with risks but this report, richly elaborated by examples and selected institutional responses, begins to lay out a roadmap based on foundations of building trust, understanding incentives, and developing appropriate business models.

The ultimate impact of the vision, principles, and roadmap sketched here, however, are only valuable if they are acted upon, and the report also calls for a disciplined and pragmatic path to progress built around the principles of *"thinking big but starting small"*, use cases, and prioritizing action over process.

If this report marks the start of a significant journey for the global agricultural community then we would do well to pay heed to the African proverb inscribed at the cornerstone of the BMGF headquarters, *"If you want to travel quickly - go alone; if you want to travel far - go together"*

Stanley Wood
Agricultural Development
Bill and Melinda Gates Foundation

TABLE OF CONTENTS

AUTHORS: Dean Allemang, Bobbin Teegarden (Working Ontologist LLC)

WITH THANKS TO: Florentin Albu (Rothamsted), Tim Davies (Practical Participation), Medha Devare (CGIAR), Lori Finch (USDA), Elisabeth Fischer (Syngenta), Max Gilmore (ANZ), Sander Janssen (Wageningen University), Johannes Keizer (FAO), Nikos Manousellis (AgroKnow), Graham Mullier (Syngenta), Cyndi Parr (USDA), Martin Parr (GODAN Secretariat), Valeria Pesce (FAO), Ben Schaap (Wageningen University), Derek Scuffell (Syngenta), Jeni Tennison (the ODI), Richard Tiffin (Agrimetrics), Paul Wester (USDA), Stan Wood (Gates Foundation)

RESPONSE PAPERS: Medha Devare (CGIAR), Silver Oliver (Agrimetrics), Jeni Tennison (ODI), Nikos Manousellis (AgroKnow)

EDITING: Anna Scott (ODI), Paul Nagle (CABI)

DESIGN: Anja Pircher Design (anjapircher.com)

August 2016

EXECUTIVE SUMMARY

Agriculture would benefit hugely from a common data ecosystem. Produced and used by diverse stakeholders, from smallholders to multinational conglomerates, a shared global data space would help build the infrastructures that will propel the industry forward.

In light of growing concern that there was no single entity that could make the industry-wide change needed to acquire and manage the necessary data, this paper was commissioned by Syngenta with GODAN's assistance to catalyse consensus around what form a global data ecosystem might take, how it could bring value to key players, what cultural changes might be needed to make it a reality and finally what technology might be needed to support it.

This paper looks at the challenges and principles that must be addressed in in building a global data ecosystem for agriculture. These begin with building incentives and trust – amongst both data providers and consumers – in sharing, opening and using data. Key to achieving this will be developing a broad awareness of, and making efforts to improve, data quality, provenance, timeliness and accessibility. We set out the key global standards and data publishing principles that can be followed in supporting this, including the 'Five stars of open data' and the 'FAIR principles' and offer several recommendations for stakeholders in the industry to follow.

- **Finding business models** that provide incentives for various entities to collect and share data. If these models provide business value directly to the data providers, the quality of the collected data will be higher.
- **Leading by example by** providing open data sources. Syngenta has already done this by publishing data about the results of its Good Growth Plan.
- **Encouraging data standards** that make it easier to produce and share data. In doing so, stakeholders will need to have reasonable expectations of how these standards will be used.
- **Automating data collection.** Automatically collected data is more likely to be accurate and precise than data collected by hand.
- **Annotating datasets.** Even automatically collected data cannot be used if it is not described in a consistent and understandable way.
- **Following data sharing principles.** The five-star maturity model and the FAIR principles provide guidelines for creating and sharing data.
- **Using the data.** All of the best data sharing efforts have little impact if the data is not used in a productive way. Stakeholders must encourage a cottage industry of data-backed apps that get the most value from datasets.

In response to the paper, four key organisations engaged in developing data infrastructures (CGIAR, Agrimetrics, ODI and AgroKnow) state what a common data ecosystem would mean to them.

INTRODUCTION

Effective and efficient use of data has transformed numerous sectors, improving the ways products are produced, distributed, utilised and developed. Advances in data innovation have surpassed all expectations; data scientists continue to be amazed at how powerful massive amounts of data can be in solving supposedly intractable problems in science, engineering, commerce and industry[1].

However, in the agriculture sector, amassing varied and pertinent datasets to reap the kinds of benefits seen in other industries is challenging, but crucial to solving systemic problems.

The world population is growing rapidly and we need agricultural productivity to keep up. In recent years, large-scale open and shared-data projects – and research on what makes them succeed and fail – have provided good examples of how to plan and conduct global data-sharing initiatives.

This paper outlines priorities for creating an effective global data ecosystem for agriculture, from engaging stakeholders to sourcing, sharing and collaborating with data. Some of these priority areas are specific to agriculture, while others are lessons learned from data sharing initiatives in other sectors. The authors draw on their experience with the Open PHACTS initiative in particular.

[1] Alon Halevy, Peter Norvig, Fernando Pereira, "The Unreasonable Effectiveness of Data", IEEE Intelligent Systems, vol.24, no. 2, pp. 8-12, March/April 2009, doi:10.1109/MIS.2009.36 [ONLINE] Available at: http://static.googleusercontent.com/media/research.google.com/en//pubs/archive/35179.pdf.

HOW CAN DATA ADDRESS GLOBAL CHALLENGES?

Feeding a growing global population will mean tackling a wide range of challenges: limiting food waste, improving resource use efficiencies, dealing with and preventing soil degradation, and coping with water shortages, to name a few - and all in a context of severe constraints imposed by weather uncertainties due to climate change. How can data help?

A wide variety of datasets underpin agriculture products and processes, which vary in size and subject matter and in how they are updated and governed. Some of the many ways that different data can be used to increase productivity in agriculture are highlighted in a 2015 discussion paper, produced by the ODI[2], and they mirror to a large extent the benefits of data analysis in other industries:

- **Geological, satellite, soil, weather and market data** are used for early, accurate detection and prediction of problems (pest outbreaks and resistance, water shortages or floods, low yields)
- **Data on best practices, weather, markets; scientific research (i.e. agronomy etc.); and longitudinal studies** are used in planning what to grow, what treatment to apply, when to plant, treat or harvest
- **Weather, crop yield, pest outbreak and production history data** are used for risk management (hedging yields, insurance) and damage control (drought, pests)
- **Regional yield variation and climate prediction data** are used for managing subsidies (funding history, financials)
- **Product and supply-chain data** are used to inform consumers (individuals as well as companies)
- **Targeted data on new and emerging pests** (e.g. Soy Rust) enables fast responses to the challenges above, and has market value.

These use cases indicate the sorts of data that can provide value when opened or shared, along with the types of stakeholders that could be willing and able to share it:

- **Commercial organizations** (product, sustainability, and yield data)
- **Governments** (satellite data, weather data, market data)
- **Traders and value-added resellers** (supply chain data, market data)
- **Science community, labs, clinics, universities** (agronomics, chemistry)
- **Farmers** (pest and yield data from precision farming[3], soil data)

In contrast to other industries, agriculture has a relatively broad spectrum of stakeholders. Agricultural product producers range from part-time, small-scale farmers to high-tech multinational conglomerates, and supply chains begin in some of the world's most remote areas. From 'farm to fork', products go from being raw materials, through processing, trading, hedging and brokering to eventually make it to the customer's table. All of the stakeholders involved in these processes are potential producers as well as consumers of data.

This paper considers ways in which data infrastructure can be built for agriculture by focusing on four key areas:

Stakeholder engagement. In order for stakeholders in agriculture to feel confident and motivated in how they share and access data, issues like data quality, curation, licensing and sustainability need to be addressed.

Data sourcing and handling. Different datasets need to be treated in different ways, depending on technical aspects, such as their size and how frequently they are updated. Data and metadata (data about data) play different roles in the data ecosystem, and must be treated accordingly.

Sharing frameworks. Many data-sharing and open data frameworks exist to help guide data exchange and use, based on data initiatives that have been established and developed over time.

Collaboration frameworks. The World Wide Web is the largest, most successful distributed data system ever built, but even so, it is not the only way to think about collaborating on the web. Other systems have had considerable success. An effective data-sharing policy will have to consider these other collaboration paradigms.

STAKEHOLDER ENGAGEMENT

Factors affecting stakeholder participation by data suppliers, users, and re-publishers largely fall into two categories: their *motivation* to participate in the first place, and *trust* that doing so will bring about good outcomes.

Building motivation for data sharing

Many groups in the agriculture sector might not have obvious motivation to participate in data sharing and use. A private enterprise, scientist or farmer might not see a reason to share data in the first place, let alone know who to share it with and under what conditions. However, the motivations for sharing and opening data can be profound once stakeholders understand the benefits, which vary widely depending on who the stakeholders are.

In the 'Better Cotton' example (as shown in the Sidebar: How Data Improves Efficiency of Cotton Production), we see a variety of stakeholders: farmers, traders, governments and brands such as Levi's and H&M.

The name brands needed the data to satisfy sustainability goals (they want to continue to source high-quality cotton for years to come) as well as marketing goals (Levi's maintains a reputation for being an environmentally conscious company); they were the main consumers of the data. The business question they wanted to answer had to do with the ultimate source of the cotton they use in their products: "where does the cotton in this pair of jeans come from?" Data from the whole supply chain is needed to satisfy this goal, from the point of production of the cotton through to each point at which the cotton was transformed (gin, spinning, weaving, dyeing, cutting and sewing). A major problem is how to motivate each of the businesses in the value chain to provide quality data. The fact that the name brands want the data motivates all of their suppliers to provide data (they want to keep their customer happy), but is frequently no direct reward for any data supplier to provide higher quality data. In order to get trustworthy data, there has to be a direct reward to the data supplier – an answer to the question, "what's in it for me?" This is a common problem in data sharing: it must be worth the data provider's effort to provide quality data. Until this issue is resolved, it will be difficult to trust the data received from intermediate stakeholders.

A large part of the motivation for data sharing has to do with how widely it will be shared, with whom and under what conditions. The ODI has developed 'The Data Spectrum' to provide a framework for understanding the language of data, from 'closed' to 'shared' to 'open'. As the dynamics of data sharing on a global scale are examined in this paper, it becomes clearer how the placement of a particular dataset or data management initiative at a particular point on the Data Spectrum has a great impact on important features for data sharing such as incentives, risks, and utilisation of data.

In many corporate settings, data held by businesses is closed and only shared on a 'need to know' basis (see Figure 1). Product development efforts are protected by trade secret in advance of being patented. The

[2] ODI. 2015. How can we improve agriculture, food and nutrition with open data? [ONLINE] Available at: http://www.godan.info/wp-content/uploads/2015/04/ODI-GODAN-paper-27-05-20152.pdf.

[3] Market Clarity/Shara Evans. ACFR: Robots Set to Transform the Automotive and Agricultural Industries. [ONLINE] Available at: http://marketclarity.com.au/acfr-robots-set-to-transform-the-automotive-and-agricultural-industries.

How Data Improves Efficiency of Cotton Production

Cotton is widely regarded as one of the more environmentally wasteful crops. Cotton fields routinely use more water, fertilizer and herbicides than many other crops. The Better Cotton Initiative (BCI) determined that this wastefulness largely resulted from poor cotton farming practices in many parts of the world. The prevalence of these poor practices has many causes, including misinformation, the transitory nature of many of the stakeholders (some cotton gins only operate for part of the year), and lack of local infrastructure to control production practices. The BCI put into place a rating program which rewarded farms for improving their practice. In contrast to certifications like Organic or Fair Trade, BCI certification was a measure of improvement, rather than a comparison to a fixed standard of good practice.

The BCI instituted a comprehensive set of programs for education, information and support for local farmers, but a key challenge was to provide an incentive for the farms to actually improve their practices. A farmer could attend a meeting where information was presented, but how do we get them to actually change how they perform?

The BCI worked out a set of incentives based on issuing BCI certificates to farms that passed an audit showing that they had in fact improved their practices. Among the members of the BCI were name brands (Levi's, H&M, Marks & Spencer, etc) with a particular stake in the future of the cotton industry. These brands set goals for utilisation of certified BCI cotton. One BCI member, Levi Strauss & Co., realised that there was a fly in the ointment. A commitment at the brand level to a particular percentage use of "Better Cotton" would only incentivize growers if there was a way to track the use of Better Cotton throughout the production chain from raw cotton to processed cotton to yarn to cloth to clothing. Levi's has a direct business relationship with only the last step in this chain; how can they be sure that Better Cotton is being tracked?

In 2012, Levi's sponsored a data-intensive project to track the provenance of the cotton in their products. While the technical challenges for linking data from production to the retail store were well within reach, the logistic challenges around the data were paramount. What data should be collected, and from which stakeholder(s)? The commitment from Levi's was sufficient motivation for their suppliers to report data, but exactly what data should they report? There are many relevant measurements, including shipping/receiving data, purchase/sale data and factory floor data. What happens if we mix these sorts of data? There is also the issue of the accuracy or currency of the data. The supplier may be motivated to provide data, but are they motivated to make sure it is correct? Collecting instrumented data helps this situation, but which parts of the process are instrumented?

All of these issues come down to how well we can trust this data. If we don't know exactly what kind of data we have collected, we cannot productively combine it to draw sound conclusions. If we can't trust the validity or currency of the data to begin with, we cannot interpret the conclusions that we draw from the data.

experiments that chemical firms (including pharmaceutical companies and pesticide manufacturers) perform during development are overwhelmingly considered proprietary.

The Open PHACTS project (see Sidebar: Open PHACTS) explores the notion of pre-competitive research: particularly research about the human genome and biochemical factors related to drug pathways that are of interest to all pharmaceutical companies. This is an example of data that are shared further to the right of the data spectrum, i.e, more openly. Some of the data result from publicly funded research, and are fully open in the sense of Figure 1. Other data were obtained through private contract, which involves some limitations, putting such data at various places in the 'Shared' section.

Scientific results (such as the pharmaceutical data in Open PHACTS and agronomic data) are typically based on data collected as part of carefully planned experimentation. The raw data from these experiments are, for a number of reasons, rarely shared in their own right. In some settings, there are few incentives for sharing intermediate data, since academic rewards are still primarily given for results reported, not the data used. But even as that changes, with data-sharing policies from funders such as the Gates Foundation, the data management skills needed to make this data available and usable are not yet common. Syngenta, through its publication of Good Growth Plan progress data (see Sidebar, Syngenta's Good Growth Plan, later in this article), is leading by example, showing how data - in this case, the progress claims for the Good Growth Plan - can have greater beneficial impact if published openly.

Enterprises often have a concern that sharing or opening data could undermine their competitive advantage. In these cases, questions of data ownership and control are key: if I share my data, do I still own it? In the case of data collected from smallholder farmers, whose data is it? Does it belong to each smallholder, or to the aggregator who collects it? If I own it, does that mean I control who can see it, and how they can use it? If I share my data today, can I retract access to it in the future? Appropriate guarantees for the control of shared data can contribute to a willingness to share it in the first place. All of these issues must be considered when making a business case for data sharing. The ODI has explored issues and benefits of data sharing within business contexts in its paper on how to make the business case for open data[5].

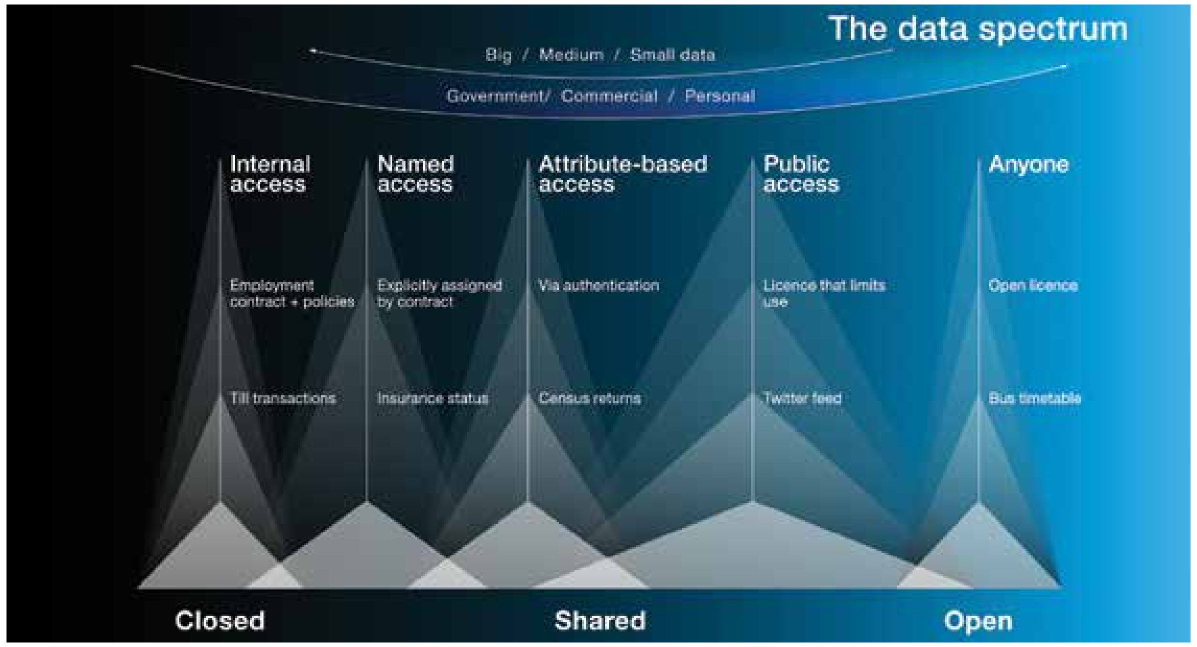


Figure 1: The Data Spectrum

Open PHACTS

Open PHACTS was a five-year project sponsored by the Innovative Medicines Initiative in the European Commission, which completed in February 2016, transitioning its operations to the Open PHACTS Foundation, a not for profit organisation whose charter is to continue to support the availability of the Open PHACTS Discovery Platform. Its charter was to make available valuable scientific data that can benefit the pharmaceutical industry as a whole. Contributors to Open PHACTS came from research labs, universities, and small enterprises, with matching effort from large pharmaceutical companies. The philosophy behind the effort was that for certain basic research, sharing the burden of collecting, curating and connecting the data would provide more value to the industry as a whole than separate efforts could provide to any single member. To some extent, the issue becomes an ethical one; by sharing research, a cure for some dire medical condition could be discovered much earlier than it could in a purely competitive space. If even a single person dies from that condition because a cure was delayed, to what extent is the industry responsible?

While the scope of the data that is managed by Open PHACTS is much more limited than what we are discussing in this paper, many of the lessons learned by the Open PHACTS project are relevant to any global data sharing initiative.

- **Linking.** Multiple datasets are valuable; having them available in a single place, even more so. Having them connect together (Berners-Lee's 'five star' data) makes them even more valuable still. One of the most beneficial outcomes of the Open PHACTS effort was the creation of so-called "link sets" – data about equivalences between entities described in separate datasets. For example, one dataset might refer to a product called "Roundup", while another refers to "Glyphosate". The equivalence of these two chemicals is maintained in a separately governed link set.
- **Data Quality.** Data collected from multiple sources around the world will be of varying quality. Even in the Open PHACTS domain, where the data had undergone some sort of peer-review process, data quality still had to be monitored by the project. This issue is so important, that a spin-off effort called Elixir[4] was created with this as a primary aim.
- **Availability.** Open PHACTS provided a wide variety of ways to access the data, from programming interfaces (APIs), to specific apps that solve particular problems. One of the major successes of Open PHACTS was that it spawned a sort of cottage industry of apps that exploit the data.
- **Sustainability.** Many of the apps based on Open PHACTS demonstrated their value, and came to be essential to various business activities. This called into question the sustainability of the data. The data service is there today – will it be there tomorrow? I can make a snapshot of the data, but much of its value is in its currency, so a snapshot will not do. How can the data services be sustained?

We wish to acknowledge the Open PHACTS Foundation, the charitable organisation responsible for the Open PHACTS Discovery Platform, without which this work would not have been possible.

[4] Crosswell LC, Thornton JM., ELIXIR: a distributed infrastructure for European biological data. In Trends Biotechnol Volume 30 (2012) p.241-242 DOI: 10.1016/j.tibtech.2012.02.002.

Building trust in data sharing

Effective data sharing depends on a strong network of trust between data providers and consumers. Infrastructure for data sharing will not be used if the parties who provide and use the data don't trust the infrastructure or one another.

For **data consumers**, trust in data sharing depends on numerous factors:

- **Knowing the source.** Trust in data begins with knowledge of its source.
- **Trusting the source.** If you know that data comes from a trusted source, then you can rely on it, and on the conclusions you draw from it.
- **Timeliness of the data.** Even when from a trusted source, data is not useful if it is outdated.
- **Data quality.** Trusted data must accurately and precisely reflect what it measures.
- **Sustainability**. A trusted dataset must have some guarantee of availability.
- **Discoverability**. Like documents, data is only useful if it is straightforward to find.
- **Documentation and support.** Consumers should be able to access support for data if needed.
- **Interaction.** Consumers should be able to provide feedback if there is a problem with data.

Trust issues of this sort are accommodated in the ODI's Open Data Certificates[6].

Data providers also have to be able to trust not only the sharing infrastructure, but the consumers of their data. Concerns of data providers might have include the following:

- Will the data I provide fall into the "wrong hands" (however that is defined)?
- Will someone use the data against me, or to support a competitor?
- Will I be held liable if someone abuses the data?
- Will someone else be able to claim the credit for findings based on the data I collected?

In constructing a data ecosystem for agriculture, stakeholders must build trust among themselves. The providers of data must be motivated to release it, and that will be undermined if the risks outweigh the expected benefits of sharing or opening it.

Building trust

A key factor in building trust in agriculture data sharing is to understand the issues faced by the wide variety of associated stakeholders. We can begin by categorising the issues that contribute to trust (or lack thereof) for data producers and consumers in the industry.

The Open Ag Data Alliance has already approached some of these challenge areas and has started to define principles for data ownership, reference implementations to test these principles, and advocate an open and standards-based approach[7].

[5] ODI. 2014. How to make a business case for open data. [ONLINE] Available at: https://theodi.org/guides/how-make-business-case-open-data

[6] ODI/Open data certificate. What You Need. [ONLINE] Available at: https://certificates.theodi.org/en/about/whatyouneed

[7] OADA. 2016. OADA Welcomes Autumn – IoT Platform. [ONLINE] Available at: http://openag.io/blog

DATA SOURCING AND HANDLING: FOR THE DATA CONSUMER

Provenance

‘Provenance’ refers to tracking the source of something. In agriculture, there is often a need to track the source of the physical materials that make up a final product (see above Sidebar: How Data Improves Efficiency of Cotton Production). This can be challenging when products undergo a chain of transformations and pass through many hands on their way to the final consumer.

Provenance in this context can also refer to the source of data. Where did a particular data set originate? When and how was it collected? Good data provenance can increase consumer confidence. Just as physical goods can undergo transformations from producer to consumer, so is the case with data. Results can be derived from raw data and then propagated as datasets in their own right. Confidence in the derived data should depend on confidence in the source data.

Taken on its own, open data, like any information on the web, has to be treated with skepticism. Just because it has been published doesn’t mean it is up to date, correct or useful. Trust in open data begins with knowledge of its source – where it came from, and who provided it. For example, self-reported data from players in a complex supply chain can vary in reliability, while weather data from an established source like the National Oceanic and Atmospheric Administration, for example, is highly curated, heavily reviewed and hence more reliable.

Trusted sources

Many datasets from well funded agencies (usually sponsored by government bodies) are already made available in standardised form. Examples of this sort of data are weather data and geopolitical data.

The scientific research and publication process can be a source of valuable data. Scientific research typically involves the collection of raw data in an experiment from which conclusions are drawn. These conclusions go through a peer-review process organised by the publishers. The results of such studies can themselves form useful open data sets (e.g., data about the genetics, metabolism and biology of various organisms, including crops and pests). Data from this process has a high degree of trustworthiness, since the reputation of the researcher and the publisher rides on the quality of the publication.

Data maintenance

Many datasets are inherently ephemeral: market data, production and consumption data, and weather data are good examples. These data are meaningless unless we know the timeframe for them and they are updated regularly. Soil data (see Sidebar: Soil Data) are similar, even if not updated as often as weather and market data. They change by location and during the course of – and between – seasons.

Soil Data

Soil data is important to many aspects of agriculture. It is a diverse discipline of itself, and includes disparate elements: lab data; field interpretation; and classification system data. In some elements it is closely allied to and draws upon schemas used in description of geospatial data - data that can be expressed in the form of points or polygons. Soil data is published by a wide variety of agencies worldwide[8]. Unfortunately different countries or research bodies use different lab standards, classification systems and even units of measurement. The creation of a single global information system for soil data is unfeasible; each stakeholder has a vested interest in the way they are currently handling their data. Even if a global standard were established, there are genuine business practices and customs that are already committed to other ways of managing data. The variety of data sources will not simply go away.

Nevertheless, there are some commonalities between datasets and these can be very useful for applications in agriculture. The challenge is to create an infrastructure that is able to formalize the commonalities between different soil datasets.

Toward this goal, the Global Soil Partnership (GSP) of the Food and Agricultural Organisation (FAO) is building a Global Soil Information System and SoilML[9], an important element for data exchange is an exchange schema for soil data similar to GeoSciML.

“Building on the new ISO 28258 for the exchange of digital soil-related data will allow owners of soil data to publish via the Web. This new standard still requires tests by a broader community of soil data holders in order to gain acceptance and to become routinely applied. This task is now also part of the work program of the Working Group Soil Information Standards (WG SIS) of the International Union of Soil Sciences (IUSS)”[10]. Standards that are used are OGC- and ISO-standards (e.g. INSPIRE in Europe).

Peer-reviewed scientific data is updated slowly in comparison, but the importance of current information is key to the industry. Having the latest scientific and agronomic results can be the difference between selecting a broad-spectrum pesticide and one that is correctly targeted to a particular invasive organism.

In order to maintain trust in data, it is necessary to have a stable policy that stakeholders will comply with about how quickly data is made available and how long it is kept. This must be balanced with the fact that keeping up with current data can pose serious technological challenges owing to the scale of the data set and the rate of its generation (see “ACFR: Robots Set to Transform the Automotive and Agricultural Industries.”[11]).

Data availability

Data production is one thing, its dissemination is another. Open data is useful when it can be delivered into the right hands (or the right machine) and within a context where it can be most valuable. In some cases, this might be a laboratory researching the efficacy of a treatment (for example, a herbicide’s effectiveness at treating some pest). Sometimes the data must be delivered into the field, so it can be used to help a smallholder make informed decisions on which crop varieties to grow or which treatments

[8] Montanarella, Luca et al. Soilml As A Global Standard For The Collation And Transfer Of Soil Data And Information. 1st ed. 2016. Web. 15 July 2016. http://eusoils.jrc.ec.europa.eu/ESDB_Archive/eusoils_docs/Poster/montanarella_EGU2010_XML.pdf.

[9] FAO. Developing SoilML as a global standard for the collation and transfer of soil data and information. [ONLINE] Available at: http://publications.jrc.ec.europa.eu/repository/handle/JRC56629.

[10] FAO. Plan of Action for Pillar Four of the Global Soil Partnership. [ONLINE] Available at: http://www.fao.org/3/az921e.

[11] Market Clarity/Shara Evans. ACFR: Robots Set to Transform the Automotive and Agricultural Industries. [ONLINE] Available at: http://marketclarity.com.au/acfr-robots-set-to-transform-the-automotive-and-agricultural-industries/.

to apply. There must be a variety of data delivery channels, fine-tuned to each case for data delivery.

The 'fine-tuning' of data delivery channels can become a business opportunity for data intermediaries, in the case where the data is fully 'open' (in the sense of Figure 1). An intermediary can provide services to customise data delivery for the vast range of customers that might exist for the data. Open data creates the possibility of a marketplace, where alternative sources of relevant data are available.

To be made available, data has to be stored in a way that makes it accessible. Even in the modern days of cloud deployment, the data and applications are stored on some hardware somewhere, even if it is virtualised. A strategy for sharing data on a global scale must specify where it will be stored, and what service level agreements (SLAs) will be maintained (up time, throughput, access controls, etc). There are a number of approaches to this:

1. **Leave the responsibility of hosting to the data provider.** Many open data sources are already hosted by the original data providers. Data provided by the United States Department of Agriculture (USDA), CGIAR Centers, and the Food and Agriculture Organisation of the United Nations (FAO) are hosted by those organisations. But the success of a global open data initiative relies largely on the democratisation of data: it has to be possible to publish data on a large scale from a multitude of sources. Supply chain data is a good example: just as we cannot expect every player in a supply chain to have their own website, it is unreasonable to expect that every member in a supply chain will host their own data for the general good of the industry.
2. **Involve a maintenance organisation whose job is to host data.** This is the solution that the Open PHACTS Foundation is taking. The foundation maintains a server with appropriate SLAs that serves up the Open PHACTS data. This solution works at the scale of the Open PHACTS data (which is a highly curated dataset, and grows relatively slowly), but could be problematic for more democratised data.
3. **Have the data hosted by a private enterprise.** The value of distributing open data has been recognised not only by initiatives like GODAN and Open PHACTS, but also by corporate investors. There is already a wide range of companies – from startups to more established enterprises – with business models based around providing hosting services for open data. It is important to note, however, that this solution only resolves the issue of hosting data. It does not solve issues around data maintenance and governance. Examples operating in this space include Socrata, CKAN, Data-Press, data.world, and AgroKnow.
4. **Involve large GODAN members in hosting data.** GODAN includes over 300 members who have signed on to its statement of purpose. Some of these members might have the resources to host data themselves, providing this service as an in-kind contribution to GODAN.

A number of initiatives are underway that can provide data hosting services for global agriculture data. Specifically, for agriculture, the Agrimetrics project in the UK is developing infrastructure for data hosting[12], CGIAR is considering this for data on global agriculture and development, and several private initiatives are developing data sharing/hosting models that could be used to manage and maintain data longevity, as explained above.

Because of the heterogeneous needs of a global data ecosystem, the solution will probably be a hybrid of many approaches (e.g. data delivery to cell phones on farms, and to analytics teams in large enterprises). A way forward that fits well with current efforts is to have publicly funded efforts (like Agrimetrics and CGIAR's Big Data and ICT platform) lead the way with experimental approaches, enabling private enterprises – either small start-up or initiatives in larger companies – to carry successful models forward. They will have to support the trust principles outlined in this paper.

Finding data

Even if data is published openly for anyone to access, use and share (see Figure 1), it will not be useful if it cannot be found by applications or users. There are a number of approaches to making data findable.

- **Provide a persistent reference.** Following the FAIR principles for *Findable, Accessible, Interoperable and Reusable*[13] data, data should be made available in a consistent location that can be referenced on the web. We expect this from ordinary web pages; datasets are no different in this regard.
- **Provide metadata for datasets.** This is a familiar method for ordinary web pages as well, allowing annotation to assist search engines in finding datasets. The World Wide Web consortium (W3C) has published the Vocabulary of Interlinked Datasets (VoID), a model for describing not only the content of datasets, but also how they link together[14], and the Data Catalog Vocabulary (DCAT)[15], a model for describing shared datasets and vocabularies on the web.
- **Index the metadata describing datasets.** With the ability to refer to datasets, indexes and search engines can use the associated metadata to enable discovery. The FAO provides the International Information System for Agricultural Science and Technology (AGRIS)[16], an index of information sources, including datasets, for agriculture. The CIARD Ring (Coherence in Information for Agricultural Research for Development) provides infrastructure for hosting and maintaining these indexes.

Efforts like AGRIS and CIARD involve curation and maintenance of indexes. For massive, on-site data – as collected in 'ACFR: Robots Set to Transform the Automotive and Agricultural Industries'[17], for example – some automated indexing of the data will be necessary.

[12] Richard Tiffin, interviewed for this paper.

[13] FORCE11. 2014. FAIR Guiding Principles. [ONLINE] Available at: https://www.force11.org/node/6062.

[14] Alexander, Cyganiak, Hausenblas and Zhao. 2011. Describing Linked Datasets with the VoID Vocabulary. [ONLINE] Available at: https://www.w3.org/TR/void/.

[15] W3C. 2014. Data Catalog Vocabulary (DCAT). [ONLINE] Available at: https://www.w3.org/TR/vocab-dcat/.

[16] AGRIS. About AGRIS. [ONLINE] Available at: http://agris.fao.org/content/about.

[17] Market Clarity/Shara Evans. ACFR: Robots Set to Transform the Automotive and Agricultural Industries. [ONLINE] Available at: http://marketclarity.com.au/acfr-robots-set-to-transform-the-automotive-and-agricultural-industries/.

DATA SOURCING AND HANDLING: FOR THE DATA PROVIDER

Licensing

A fundamental problem for any open data movement is how to deal with data licensing and governance. While it is possible for a large number of stakeholders to join in the vision of GODAN, when it comes to actually publishing data openly, most data will come with some strings attached. Even open licences like the Creative Commons CC-BY and CC-BY-SA[18] make specifications for how the data can be used. It is difficult to determine what licence should apply to results that make use of data governed by multiple licences[19].

Other data-sharing initiatives have faced this problem and approached it in a number of ways, with varying degrees of success:

- Using a no-tech solution, where data users are informed of the licences associated with all the data in a system and themselves left to determine whether they are in violation (this is the approach taken by Open PHACTS (see earlier Sidebar: Open PHACTS)
- Enforcing a blanket (open) policy across an entire portal
- Providing a constrained sets of compatible licences that can be used within the portal (e.g. limiting to CC licences, but providing a choice from several CC licences)
- Expressing licensing constraints using a machine-readable vocabulary e.g. Creative Commons[20] or the Open Digital Rights Language (ODRL)[21].

A more comprehensive approach would be for the data system to determine what conclusions can be reached according to any particular licence (including commercial licences). If some data is protected by a commercial licence, then it may be possible for a data platform to manage licence keys and only provide the information permissible under the licences. This would involve extra technical infrastructure that does not yet exist, but could be developed by a GODAN partner, or a third party (e.g. a startup in the open data hosting space) could be convinced to provide this as a featured service.

Security

If sensitive data is to be shared, there must be provisions in the platform to ensure security of that data. Whether data is closed or shared with specific individuals or organisations, it will need to be hosted in a controlled way. Depending on the sensitivity of the data, this will include some guarantee of security, e.g. against hacking. In the most extreme cases, the security requirements for shared data in agriculture could be as severe as for shared data in the military.

These principles are not unique to agricultural data, and have been studied in depth[22]. The basic concepts behind these principles are that services should be hard to compromise, that a compromise should be easy to detect, and that the impact of a compromise can be contained.

For open data (on the right-hand side of Figure 1), this is much less of a concern, but to build trust among data providers, some support for data security must be in place.

Dataset issues

Managing data on a global scale involves a wide variety of data types. The stakeholder engagement issues already discussed (trust and motivation) can take very different forms, depending on the data types under consideration. Technical issues also vary depending on dataset type, in terms of how data is governed, its technical infrastructure, and the particulars of collaboration approaches.

To illustrate the ways in which datasets can differ, some examples that have been used in data-sharing projects are set out below:

1. **Scientific data** including biology, chemistry, meteorology, agronomy and related fields. For example, the Open PHACTS project (see Sidebar: Open PHACTS) curates and distributes data about biological systems, in particular data that applies to the testing and development of drugs. This sort of data is highly curated (typically through peer-reviewed journals), and the result of extensive research. The use of a single data point can become part of a drug-discovery process worth billions of dollars. While the Open PHACTS data (especially in biochemistry) is applicable to agriculture, there is a wide range of scientific information about crops, invasive organisms, soil chemistry or toxicity, for example, which could pose similar challenges.
2. **Production, consumption, purchase and sale data.** The Better Cotton Initiative (see Sidebar: How Data Improves Efficiency of Cotton Production) tracks the provenance of cotton throughout a global supply chain. This involves information about production and consumption of raw materials, and how they were shipped and sold. By combining this information, conclusions can be drawn about the origins and destination of the cotton.
3. **Precision instrumentation data.** With the use of ever advanced machinery to perform the basic functions in agriculture[23] [24], the opportunity arises to track detailed information about the state and treatment of soil, invasive species, crop health and other important factors. The use of airborne drones to monitor agricultural factors is also becoming more feasible.
4. **Satellite and weather data.** The National Oceanic and Atmospheric Administration (NOAA) in the USA and the Meteorological Office (Met Office) in the UK, as well as other agencies around the globe monitor weather data on a regular basis.

   Advances in image processing make satellite data more effective. Satellite data availability can be unreliable, owing to cloud cover and other weather conditions, but it can still cover areas that are otherwise unreachable due to political or geographical limitations.

Data size and speed

For data sets that never change, it is sensible to talk about the size of the data set and the storage space it would need. However, many data sets are collected continuously in real time, and for these data sets, size is not a meaningful measure, since the set is always growing. For such data sets, a better measure is speed - how quickly does the data set grow.

For some types of instrumentation (sequencers, remote sensors), the speed at which data is collected is so high that it is difficult to get the data to a storage facility[25].

The size/speed of a dataset is a key consideration when it comes to sharing or archiving it. In the case of Open PHACTS, the speed is fairly low, since new scientific data is reviewed rather slowly. But scientific data encompasses a wide range of topics and can be quite large. The Open PHACTS dataset is largely made up of verified facts about the reactions of certain substances and steps in various processes. It is conceivable to have all of the Open PHACTS data running in a single database instance all at one time. Datasets of this sort are often measured in terms of how many facts (sometimes expressed as "nano-publications" or as "triples") are included.

In contrast, production and sales data usually record specific information about particular transfers. The data recorded often comes from ledger sheets or purchase orders, and hence is limited in scope. This limits the effective speed of growth of the dataset. But since longitudinal purchase data is important for tracking trends, the size of the relevant data is unbounded.

[18] ODI. 2016. Reuser's Guide to Open Data Licensing. [ONLINE] Available at: https://theodi.org/guides/reusers-guide-open-data-licensing.

[19] The Open Data Institute. 2013. Licence Compatibility. [ONLINE] Available at: https://github.com/theodi/open-data-licensing/blob/master/guides/licence-compatibility.md.

[20] Park, Jane. "NSF Grantee Opens Up Wind Technology Training Materials For Fellow Grantees - Creative Commons". Creative Commons. N.p., 2014. Web. 15 July 2016. https://creativecommons.org/2014/05/28/nsf-grantee-opens-up-wind-technology-training-materials-for-fellow-grantees/.

[21] "ODRL Version 2.1 Core Model | ODRL Community Group". W3.org. N.p., 2016. Web. 15 July 2016.

[22] CESG. 2016. Security Design Principles for Digital Services. [ONLINE] Available at: https://www.cesg.gov.uk/guidance/security-design-principles-digital-services-0.

[23] Market Clarity/Shara Evans. ACFR: Robots Set to Transform the Automotive and Agricultural Industries. [ONLINE] Available at: http://marketclarity.com.au/acfr-robots-set-to-transform-the-automotive-and-agricultural-industries/.

[24] The Technical Centre for Agricultural and Rural Cooperation (CTA), Data Revolution for Agriculture. 2016.

[25] Market Clarity/Shara Evans. ACFR: Robots Set to Transform the Automotive and Agricultural Industries. [ONLINE] Available at: http://marketclarity.com.au/acfr-robots-set-to-transform-the-automotive-and-agricultural-industries/.

Precision instrumentation records data about particular values measured at certain locations and times. This is not limited by a traditional, paper-based system, and can include a broad range of values. Storage of this data can even be limited by the speed of the connection to a data collection site[26]. While this sort of data can also be seen as a set of facts, usually such datasets are measured in terms of gigabytes, to decipher how much space would be needed to store them, or how much bandwidth to transmit them.

Satellite data and weather data are typically managed by large agencies, as the infrastructure needed to record them is usually only available to governments or similar large entities. Satellite data can have a significant time lag, due to the slower speed of data being collected and made available, and the limited bandwidth of transmitting the data back to a ground station.

Satellite data in particular is considered ‘heavy’, that is, the bottleneck in its usefulness lies in the difficulty in moving the data from one place to another. The ‘heaviness’ of satellite data is one factor that keeps agriculture applications from using the highest-resolution data available today.

Data structure

Perhaps the most striking difference between satellite data and other types of data is its structure (or rather, lack thereof). Instrumented datasets are largely tabular: each reading consists of the same measurements, varying by location and time. Satellite data, in contrast, consists typically of image data, which can be massive; these images can be processed by sophisticated algorithms to detect the location of objects on the ground (e.g. 1 square metre per image pixel). The market for exploiting satellite data (e.g. by improved image processing) is competitive, and several companies (like Orbital Insight and The Climate Corporation) have found niche value from this sort of data.

Precision farming instrumentation includes specific, structured measures, but new work includes photographic data taken from airplanes or drones. These datasets have the same unstructured properties as satellite imagery, but without the space-to-earth limitations.

Scientific data such as that curated by Open PHACTS is highly structured, but given the variety of data, there is a large variation in the structure. Much of the data (about metabolic pathways, chemical structure, and processes, for example) is best represented as graphs, while other data (composition, weight, etc) is best represented in tabular form.

Data vs. metadata

A distinction that is often cited when dealing with data management is that of data vs. metadata (i.e. data about data). There are a number of specific distinctions that these might refer to:

- **Metadata as schema.** When we collect tabular data, we need to know what the ‘columns’ in the data refer to. Even in non-tabular data, some schema information is helpful for interpreting the data. Many data formats include ways to specify schema metadata, e.g. XSD for XML, RDFS for RDF, and DDL for databases.
- **Bibliographic metadata.** Librarians and library scientists have used metadata to describe documents (books, articles, pictures, etc) for centuries, and have determined structures for recording and searching this kind of metadata. This sort of metadata includes provenance information (authorship, publication data), dates, size of the publication (e.g. in page count), and is applicable to datasets as well (e.g. DCAT[27] and VoID[28])
- **Shared vocabulary.** Alignment of different datasets is a challenge in any distributed setting. A key tool in governing such datasets is the use of a shared vocabulary. The vocabulary is used in the content of the data, rather than describing the data per se. For example, the AGROVOC (for AGRiculture VOCabulary)[29] provides (amongst other things) terminology for talking about agricultural products, e.g. milk, milk byproducts, milk fat, etc. Agroportal[30] and the VEST registry[31] provide access to many vocabularies related to agriculture.

Metadata is much smaller than the data it describes, but is key to the combination of datasets in a meaningful way.

For tabular data – such as instrumental data and weather data – schema metadata can be implicit in the measurement method. Open PHACTS includes a good deal of its schema metadata (which changes frequently) in its publication, since there is such a wide variety of data that it describes. Much of the data in Open PHACTS is itself vocabulary metadata, providing a way to identify a particular drug or compound so that reference to any data is unambiguous. For datasets of this sort, identity is key: exactly what drug are we talking about? Open PHACTS has a sophisticated method of specifying in exactly what context an identifier should be interpreted.

Satellite imagery involves mostly bibliographic metadata, since the data itself is not structured. This metadata provides information for interpreting the data as an image, so that algorithms can work over it.

[26] Market Clarity/Shara Evans. ACFR: Robots Set to Transform the Automotive and Agricultural Industries. [ONLINE] Available at: http://marketclarity.com.au/acfr-robots-set-to-transform-the-automotive-and-agricultural-industries/.

[27] W3C. 2014. Data Catalog Vocabulary (DCAT). [ONLINE] Available at: https://www.w3.org/TR/vocab-dcat/.

[28] Alexander, Cyganiak, Hausenblas and Zhao. 2011. Describing Linked Datasets with the VoID Vocabulary. [ONLINE] Available at: https://www.w3.org/TR/void/.

[29] FAO. 2016. AGROVOC Multilingual agricultural thesaurus. [ONLINE] Available at: http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus.

[30] “Welcome To The IBC Agroportal | IBC Agroportal”. Agroportal.lirmm.fr. N.p., 2016. Web. 15 July 2016.

[31] “VEST Directory | Agricultural Information Management Standards (AIMS)”. http://aims.fao.org/vest-registry. N.p., 1980. Web. 15 July 2016.

SHARING FRAMEWORKS

Many of the considerations outlined above have been incorporated into data sharing frameworks and maturity models. To a large extent, these efforts build upon one another, offering different guidelines with which data providers can publish more effectively and with greater impacts.

Five-star open data

Sir Tim Berners-Lee, the inventor of the web, has identified a maturity model for data publication on a global scale[32] [33]. The five-star maturity model is particularly oriented toward describing structured data and towards distributed publication of data, both of these aims relying on a strong notion of identity to make the data coherent. The model has the following elements:

★ Make data available

Making data available (and being specific about how it is licensed (see Figure 1)) is a seemingly small step but is fundamental: it means data providers are incentivised to share data, and how it is to be shared in terms of its licensing is already determined. This level of sharing already imposes a technical challenge, in that the data must be stored somewhere and access has to be publicised. However, it does not guarantee any particular utility to the data; the accuracy, provenance, currency and context are not included in one-star data. Often, data at this level is only readable by humans (i.e. in images or other documents not easily read by machines).

★★ Make the data structured (machine readable)

The next maturity level is to structure the data enough that it can be read by machine. This includes formats that require specific, often proprietary software to read. Microsoft Excel is a common example at this level; anyone who wants to read Excel data has to enter into a business relationship with Microsoft. This has an impact on data longevity – what happens to the data if that proprietary format ceases to be supported? Conversion from one proprietary system to another can be problematic, since detailed information about proprietary formats is typically not available.

★★★ Ensure data is in non-proprietary format

By replacing a proprietary format with a non-proprietary one, the availability of the data is no longer beholden to a particular commercial entity. This allows anyone in a global community to access the data. It also enables longevity, since anyone can pick up the support of a non-proprietary format, or convert it to a new format.

★★★★ Make data referenceable

This is the key to allowing the data to participate in a linked data world. It has to be possible to refer to other datasets, which makes it imperative to annotate the data with metadata – to record its provenance, accuracy, context, etc. Until datasets are well described using consistent metadata elements or schemas, it will be difficult to collect, share or interlink them.

★★★★★ Link data to other data for context

This is the grand prize for linked data, where the work is done of of actually connecting data to other sources. This is an open-ended activity, since there is potentially always another relevant data context. To 'socialise' data for global use, datasets also need to be connected to information about provenance and currency, and to shared vocabularies that describe what the data means.

The five-star maturity model guides data publishers on how to publish data to the web, and draws heavily on the web's infrastructure and governance patterns. It uses universal web identifiers (URIs) to identify resources, and links them together by making reference to other web resources. The five-star model does not insist on any particular technology, but strongly indicates the principles that defined the Resource Description Framework (RDF)[34]. It is not necessary to use the RDF to implement five-star data, but users without it may be doomed to reinvent it.

There is a challenge for reaching the level of 4-star data for large, tabular and image datasets, since a fully referenceable data format (e.g. RDF) is quite expensive for large datasets. In the worst case, this can transform 'lightweight' datasets (which can be transmitted with little demand on resources) into 'heavy' ones (where there is considerable cost associated with the transfer of the data).

FAIR principles

A number of considerations have already been outlined that need to be taken into account in creating a network of trust. Having experience in similar efforts in the pharmaceutical industry, Force11 has published the *FAIR Guiding Principles for Findable, Accessible, Interoperable and Reusable data*[35]. These principles help to ensure that the data collected and shared will continue to be usable by stakeholders in the future, helping to maintain trust in the long term. FAIR provides guidance to the creation of data standards that will support the creation and longevity of a global data ecosystem.

The FAIR principles do not specify any particular technology for publishing and managing data, but many of the technical issues underlying FAIR are achieved by using the W3C web standards of RDF and RDFS. As the FAIR overview points out, putting data on the web is not enough; appropriate treatment of licensing is also needed.

Details of FAIR are given in the cited document, but the basic principle is that data must be findable, accessible, interoperable and reusable.

- **Findable** means that there is a unique, persistent way to refer to the data. The web standard URL satisfies this requirement quite well.
- **Accessible** means that the data can be obtained by humans and machines. URLs along with the web infrastructure go a long way to satisfying this goal, but further development may be needed to handle data licensing in a useful way.
- **Interoperable** means that the data can be managed by a machine. This covers parsing standards, self-description of the data, and ways to link the data to shared resources. RDF and RDFS provide considerable capabilities along these lines.
- **Re-usable** means that in addition to the first three principles, the data is described richly enough in terms of shared data (and metadata) resources to make its application apparent.

The FAIR principles go beyond the five-stars of open data to talk about how data can be used. To get synergistic value from data, it must be available and reliable. The FAIR principles provide guidelines for how to continually improve the value of the data we produce.

The first of the principles, 'findable', poses a challenge for large-scale agricultural data. Unlike purely scientific data, a great deal of the data relevant to agriculture is very large-scale, coming from precision instrumentation or satellites. Searching through the contents of extremely large datasets of this sort is not practical. For such datasets, meaningful metadata that describes the content and context of the data is essential for its findability.

A great example of applying the FAIR principles in a particular domain is the FAIRDOM programme[36], which provides the platform and infrastructure for researchers in systems biology. This platform allows any data practitioner or owner to be in control of collecting, managing, storing and publishing their data, models, and operating procedures.

[32] Tim Berners-Lee. 2009. Linked Data ("Five Stars"). [ONLINE] Available at: https://www.w3.org/DesignIssues/LinkedData.html.

[33] Tim Berners-Lee. 2009. 5-star Open Data. [ONLINE] Available at: http://5stardata.info/en/.

[34] Semantic Web. 2014. Resource Description Framework (RDF). [ONLINE] Available at: https://www.w3.org/RDF/.

[35] FORCE11. 2014. FAIR Guiding Principles. [ONLINE] Available at: https://www.force11.org/node/6062.

[36] "About FAIRDOM | FAIRDOM". http://fair-dom.org/. N.p., 2016. Web. 15 July 2016.

Open Data Certificates

More so than any of the preceding frameworks, the ODI's Open Data Certificates[37] take into account stakeholder issues around sharing data, including issues of motivation, trust and licensing. The Open Data Certificates provide a measure of how responsive a dataset is to stakeholder issues.

One of the main aims of the Open Data Certificates is to provide assurance of datasets' sustainability. This is a lesson that was learned the hard way by Open PHACTS; success in publishing open data and creating a community of users does not automatically guarantee data's sustainability. Many open datasets are funded by grants and government projects whose duration is limited; even hosting the data after the completion of the seed grant period can be prohibitive.

Certified data includes documentation and support for the data, so there is a system for maintaining and delivering it. There are several levels of certification and the desired level depends on the goals and governance of the data provider, unlike the case of data classified as five-star or FAIR. The most highly certified data includes a support team, data provenance, and specific service-level agreements.

Licensing is built into even the first level of certification (Bronze), guaranteeing that consideration has been made for the legal use of the data, right from the start.

Silver-certified data includes support from the publisher. It is not enough to simply provide the data, it has to have some support and documentation with it. User feedback (and perhaps even a crowdsourcing policy) is included at this stage.

Gold-certified data includes machine processing for many of its support functions; machine-readable licensing, open standard publishing format (like three-star data), and an update schedule.

Platinum-certified data includes provenance information to support trust, along with unique identifiers (identification is a pervasive problem in any data publishing setting). Platinum-certified data includes the features of five-star data, with a strong support system for the sustainability of data availability.

[37] ODI/Open data certificate. What You Need. [ONLINE] Available at: https://certificates.theodi.org/en/about/whatyouneed.

COLLABORATION FRAMEWORKS

It is one thing to share data, but to achieve the desired gains from a data ecosystem for agriculture, to draw conclusions across the globe to guide decision making, it is necessary to exploit synergy between datasets efficiently.

Many of these datasets are useful in their own right, but the utility of a dataset invariably increases when it can be used in conjunction with others. For example, it is highly valuable to combine weather data with geophysical data when seeking to understand the impact of weather on a particular field where crop yields, pesticide administration or water usage is being managed. Agronomic data (e.g. soil chemistry) is valuable when it interoperates with information about organisms that are present in the soil in a real field. Connecting production data from a supplier with consumption data from its customer, and the production data of that customer to the consumption data from its customer, and so on, forms a supply chain. Combining all of these can provide a brand with insight into the original sources of its products.

Useful global data publication is a matter of data collaboration. Having already examined the stakeholder issues around trust and motivation, the particular challenges around different types of data, and some data-sharing maturity models, makes possible great progress in providing policy advice for the publication and reuse of data on a global scale.

Many of these approaches are predicated on the understanding of collaboration technology based on familiar web infrastructure, which is primarily a caveat emptor approach. Now that publication on the web is easy (and in the age of social network, becoming easier all the time), the burden of trust is on the consumer. Search engines and ratings sites gather data about web publications to guide the consumer, but in the end, it is up to the consumer to make their choice. In order to encourage and enable productive collaboration on a global scale, it is necessary to move beyond this simple model to something that supports collaboration in a sustainable way.

Systems of governance

Even in small systems, organising data from multiple stakeholders in a useful way can be challenging (see Sidebar: How Data Improves Efficiency of Cotton Production). The datasets, if they are published at all, appear in an ad hoc way, without any coordination, schema alignment or controlled vocabularies to guide the consumer about how to combine them.

The key to bringing together data from disparate sources in a productive way is some use of shared metadata, a touch point between the data sources as a common reference context. The book, "Semantic Web and the Linked Data Enterprise"[38] describes how this works within an enterprise, where shared vocabularies play a key role in providing common context between datasets. But how to establish these common vocabularies and govern their use?

At one extreme is the 'data standardisation' approach, in which some authority establishes a standard set of terminology that provides context for all datasets. As long as every data provider conforms to the standard, interoperability can proceed smoothly. While this approach is conceptually simple, it has its drawbacks, the primary one being shown in Figure 2.

The problem is that it is too easy for multiple authorities to provide guidance on the same metadata. When this happens, there is a choice to be made by any data provider, of which standard to use. Since no standard exactly meets the needs of any provider, there is continual pressure to create new standards.

Another problem that also appears in Figure 2 is that this situation happens again for each domain in which data is to be collected and connected. In the figure, the standards in question are for A/C Chargers, Character Encodings, Instant Messaging, etc. This applies equally well to standards for Chemistry, Soil, Invasive Species, Production, Sales, etc. Not only is there a proliferation of standards in any domain,

[38] Allemang, Dean, 2010, Semantic Web and the Linked Data Enterprise. In: D. Wood ed., Linking Enterprise Data. Springer, US, 2010. [ONLINE] Available at: http://linkeddatadeveloper.com/Projects/Linking-Enterprise-Data/Manuscript/led-allemang.html.

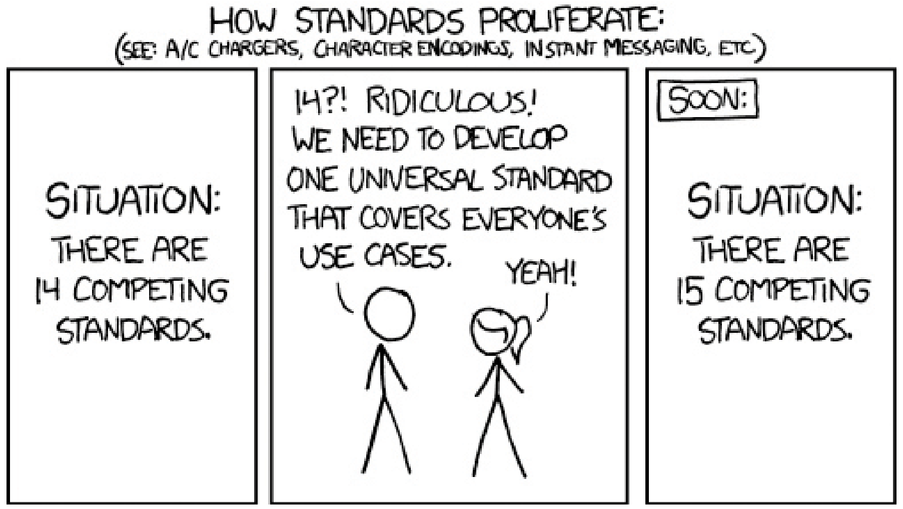

HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

Figure 2 How Standards proliferate. Source: xkcd (http://xkcd.com/927/)

but there is the issue multiple domains might be connected effectively. Agriculture is particularly challenging in this regard, because of the wide variety of relevant domains.

At the other extreme is the ‘linked data’ approach. Linked data begins with the observation that in a global setting, data will be provided by a wide range of stakeholders, and even in the presence of excellent reporting standards, it is unrealistic to expect any two datasets to necessarily share any particular standard. Hence the emphasis in linked data is not on producing any particular standard (or specifically, on producing a new standard as in Figure 2), but instead on interoperating between standards. Therefore, the linked data approach emphasises mapping between standards (as an example of mapping see the Sidebar: Open PHACTS, the success of which was a result in large part of the careful curation of linkages between compounds in various datasets).

SoilML (Soil Markup Language, Sidebar: Soil Data) is an example of how fine the dividing line can be. SoilML runs the risk of being ‘just another standard’, as shown in Figure 2. If this happens, then it will be seen simply as another requirement for data publication, and will become a barrier rather than a boon to data sharing. On the other hand, if it is promoted and developed well, it can play the role of a common vocabulary for interoperating existing datasets, without trying to supersede them. This is largely a matter of promotion and philosophy of use. The success of SoilML is due in large part to its promotion as a means of connecting datasets, rather than as a format intended to replace any others.

While data standardisation focuses on getting data providers to behave in a uniform way, linked data focuses on being able to function when they do not. As we develop in this area, some combination of these approaches will have to be followed. There is value in standardisation, but we need to be able to cope with multiple standardised approaches.

Controlled vocabularies

Controlled vocabularies play a special role in a data sharing infrastructure. They provide a way to reference a set of terms that can be used to describe information artifacts, including both documents and datasets. Controlled vocabularies can be subject to the same proliferation of standards shown in Figure 2, or they can be a cornerstone of a linked data strategy. Thanks in large part to the work of the FAO and the National Agriculture Library, the vocabularies on agriculture (Agrovoc[39] and NAL[40]) have been connected with a curated mapping between the terminologies. This is an active area of development. Mappings between these vocabularies and many others (including the CAB[41] thesaurus and the Chinese Agricultural Thesaurus[42]) are underway.

[39] FAO. 2016. AGROVOC Multilingual agricultural thesaurus. [ONLINE] Available at: http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus.

[40] USDA. 2016. National Agricultural Library Agricultural Thesaurus. [ONLINE] Available at: http://agclass.nal.usda.gov/.

[41] CABI. 2015. CAB Thesaurus. [ONLINE] Available at: http://www.cabi.org/cabthesaurus/.

[42] A. C. LIANG †, M. SINI†. Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures . [ONLINE] Available at: http://www.fao.org/3/a-ag862e.pdf.

Social networks of data (and metadata)

A strong global data policy will include data standards (shared metadata for expressing and exchanging information) as well as a linked data infrastructure (to link together datasets that were developed using contrasting or even competing standards). The goal is to move toward as much coherence as possible in the data system. Toward this end, it is helpful for a data publisher to understand what standards are being used in what community, to help them target their data with the least disruption.

This suggests a framework for sharing data that draws on social networking ideas, from the world at large (LinkedIn, Facebook) or from intramural networks (wikis, JIRA)[43]. Such frameworks have already been successful at bringing social dynamics to bear on information sharing.

Git (and Github)

Many of the social and governance aspects of information sharing have been pioneered in the open-source software space. Git[44] is a software sharing system that allows teams of software engineers to collaborate on the development of program code. Github[45] is a cloud platform that hosts Git repositories, and is very commonly used for open source code development.

Git provides services that allow teams of developers to govern the evolution of their software, including management of versions, keeping revision histories, computing differences between various versions, and resolving collisions when multiple developers make conflicting changes. While Git has been used so far primarily for computer program source code, many of the features of Git, including the governance of feedback to the corpus, are relevant to the governance of shared datasets as well.

Blockchain

A more recent development in sharing technology is blockchain. The full impact of blockchain is yet to be realised, but it is based on bringing the ideas of consensus in to the fabric of data sharing. The content of a blockchain is maintained by consensus, thereby forcing an attacker to compromise a distributed system in order to damage blockchain data. This can have deep ramifications in terms of trustworthiness of data; if it is secured in a blockchain, then its integrity and source are guaranteed. The ODI has written a paper on applying blockchain technology in global data infrastructure, looking at the landscape of use cases, privacy implications and so on, which helps gain a broader picture of the technology’s potential[46].

Culture

Collaboration platforms can be technical tools that help solve part of the data-sharing challenge. However a greater challenge to solve is the cultural inertia that prevents sharing in the first place. As long as a culture persists of not wanting to connect or share data, then no amount of technical innovations will bring about the change needed to develop a global data infrastructure.

As specified in the description of the linked data enterprise[47], it is necessary to have a community that produces and publishes data so it can be connected to other data and become more useful. Not only do the mechanics of this need to be met by technology, but a cultural change is needed that encourages information connection, so that thinking about how data can be shared and connected is part of the DNA of the data community.

[43] Wikipedia/Confluence. 2016. Confluence (software). [ONLINE] Available at: https://en.wikipedia.org/wiki/Confluence_(software).

[44] “Git”. https://git-scm.com/. N.p., 2016. Web.

[45] “Build Software Better, Together”. https://github.com/. N.p., 2016. Web.

[46] James Smith, Jeni Tennison, Peter Wells, Jamie Fawcett and Stuart Harrison. 2009. Applying blockchain technology in global data infrastructure. London: the Open Data Institute [ONLINE] Available at: http://theodi.org/technical-report-blockchain-technology-in-global-data-infrastructure.

[47] Allemang, Dean, 2010, Semantic Web and the Linked Data Enterprise. In: D. Wood ed., Linking Enterprise Data. Springer, US, 2010. [ONLINE] Available at: http://linkeddatadeveloper.com/Projects/Linking-Enterprise-Data/Manuscript/led-allemang.html

RECOMMENDATIONS FOR BUILDING A GLOBAL DATA INFRASTRUCTURE FOR AGRICULTURE

How to move from the current situation to a collaborative, shared global data space? It is clear that this cannot be achieved by a single entity, it has to be engineered through the collaboration of a variety of stakeholders working together to build the data ecosystem that will propel the whole agricultural sector forward.

There is no blueprint for building this ecosystem, but there are areas to focus on that will to allow it to develop, as discussed already:

- **Building trust.** This is the keystone for the whole infrastructure. Providers must trust that their data will be used well. Consumers must trust that data they use is correct, up to date, reliable, and available.
- **Developing standards and linking data.** Given the proliferation of standards, linked data approaches and resources must be honed so that they function in the standards "jungle". However complicated the landscape, standards can provide anchors to help interoperate and link across data to make it more meaningful.
- **Ensuring sustainability.** In order to ensure that data continues to be updated and made available, it is essential to consider data storage, transmission and backup, and governance issues (including funding) that will enable the ongoing maintenance of data.
- **Providing incentives.** The various stakeholders in agriculture must be motivated (by carrot or stick) to continue to provide timely, high-quality data.
- **Data publishing principles.** The Five-star model and the FAIR principles provide specific guidance about what it means to produce data of real value to a community. Both of these are incremental, allowing data to be published and its value refined later on.

Specific efforts will need to be made by individual enterprises, governments or initiatives, informed by the areas set out above, to develop a global data infrastructure for agriculture. These include:

- **Finding business models** that provide incentives for various entities to collect and share data. If these models provide business value directly to the data providers, the quality of the collected data will be higher.
- **Leading by example by** providing open data sources. Syngenta has already done this by publishing data about the results of its Good Growth Plan.
- **Encouraging data standards** that make it easier to produce and share data. In doing so, stakeholders will need to have reasonable expectations of how these standards will be used.
- **Automating data collection.** Automatically collected data is more likely to be accurate and precise than data collected by hand.
- **Annotating datasets.** Even automatically collected data cannot be used if it is not described in a consistent and understandable way.
- **Following data sharing principles.** The five-star maturity model and the FAIR principles provide guidelines for creating and sharing data.
- **Using the data.** All of the best data sharing efforts have little impact if the data is not used in a productive way. Stakeholders must encourage a cottage industry of data-backed apps that get the most value from datasets.

These actions will require dedication, creativity and courage. The GODAN initiative[48] has brought together hundreds of partners who have made a commitment to improving how data can help the agriculture industry feed the world. Making a commitment is not enough – creative thinking is needed to devise successful initiatives and actions that follow through to delivery. The recommendations and principles discussed in this discussion paper are intended to guide those initiatives, and to make sure that they keep us on track to keep pace with the global challenges that we face.

THE SYNGENTA CONTEXT

Following Syngenta's first publication of open data in the form of the Good Growth Plan (23 April 2015, see Sidebar, Syngenta's Good Growth Plan), the Data Sciences group in Syngenta realised it was contributing to what amounted to an immature discipline. While a lot of discussion was occurring around the sharing and use of data, especially open data, little of the debate related to agriculture and food. Data infrastructures appeared to be fragmented and there seemed little understanding of what the broader data ecosystem needed to look like for our industry. It has become increasingly clear that no single government agency or corporate entity can make the industry-wide change that is needed to acquire and manage the required data.

As a response to this realisation, Syngenta commissioned this discussion paper with the aim of catalysing some consensus around what form a global data ecosystem might take, how it could bring value to key players, what cultural changes might be needed to make it a reality and finally what technology might be needed to support it. Over the past year the idea has developed significantly through working with GODAN partners and the ODI membership, producing not only this paper but a number of responses to it from some key organisations that might benefit from a more coherent ecosystem. The responses add to the core message, building on its ideas and introducing other valuable perspectives, giving an opportunity to move beyond discussion to action. It is hoped that this set of material will catalyse action in the broader community – farmers, researchers, governments, large and small commercial organisations, and consumers – to develop the global data ecosystem which is surely needed if we are to address the significant challenges the agriculture and food sector faces.

[48] GODAN. 2011. Statement of Purpose. [ONLINE] Available at: http://www.godan.info/pages/statement-purpose.

Syngenta's Good Growth Plan

Starting in 2013, Syngenta made six commitments for contributing to sustainability of agriculture in the world. All six commitments can be stated simply:

- Make crops more efficient
- Empower smallholders
- Rescue more farmland
- Help people stay safe
- Help biodiversity flourish
- Look after every worker

Each of these commitments can be measured, in terms of crop productivity, outreach to smallholders, implementation of good soil maintenance practices, measurement of the extent of biodiversity practices, safety training attendance and labor monitoring. Rather than simply claiming that they have made progress on all of these fronts, Syngenta publishes its status data on all of these measures as open data. The data has earned a silver certificate according to the ODI certifications (see text), based on the comprehensive and objective documentation of the meaning of the data. This allows anyone who is interested in the Good Growth Plan not only to monitor its progress, but to understand how that progress is measured and to review it on that basis. The license terms of the data, the update schedule and period of coverage of the data are made clear in all datasets.

This provides a model of how data can be shared in a more general setting. From a business analysis point of view, most people are interested in the results - how much progress did Syngenta make, and how quickly. But suppose we want to know how it is measured? How do you measure smallholder empowerment, or looking after workers? Because this is Silver-certified data, anyone can review how these were measured, and exactly where and when how much progress was made. For example, In the data reported so far, empowerment is measured by outreach, and looking after workers is measured by a checklist that is published along with the data.

This is directly analogous to scientific data - scientists are rewarded for publishing results, but the way they came to those results, what they measured and the values they got for those measurements, are not usually part of the publication. We applaud publication of the data that supports these scientific results. Syngenta is leading by example, showing the data and the computation behind its claims of progress.

Good Growth Plan progress data is available for download[49].

RESPONSE PAPERS

CGIAR RESPONSE

Medha Devare

CGIAR's fifteen geographically and scientifically diverse Centers, along with other entities involved in agricultural research and development, are charged with accelerating innovation to tackle challenges at a variety of scales from the local to the global. However, the varied research outputs required for this are too often not easily discoverable, and "dark data" is common - often residing on individual laptops, not being well described, indexed, or stored to be accessible and usable by the wider scientific community. CGIAR Centers have made strong progress in implementing publication and data repositories that meet minimum interoperability standards; however, many of these still represent silos whose contents are not generally easily discoverable or inter-linked where appropriate and useful (e.g., agronomic trial data with socioeconomic or adoption data in the same geographies). In the absence of such interoperability-mediated discovery, "open" is of limited utility. The overall objective, then, is to open CGIAR's trove of research data and associated information for indexing and interlinking by a robust, demand-driven cyberinfrastructure for agriculture, ensuring that research outputs are open to enhance innovation, impact, and uptake.

There are, however, challenges to achieving this goal, among the foremost of which is that "open" is still largely an unfunded mandate, making it difficult to operationalize. Further, there is still significant scientist concern about making data open, largely centered on issues of trust, time, and quality, resulting in repositories frequently exposing metadata rather than the data sets themselves. While this qualifies as improvement, it continues to impose barriers to data access, discoverability and integration, without which complex challenges to global agriculture development cannot be effectively addressed. Echoing the Allemang and Teegarden paper, CGIAR is addressing the urgent need to create a data sharing culture and enabling environment for Open Access and Open Data (OA/OD) in parallel with the technical infrastructure mentioned above.

The technology necessary to do this exists - by harvesting content from diverse platforms, repositories, and databases, integrating different types of resources and disciplines in meaningful ways, seeing patterns in and mining data "big" and small, and seamlessly leveraging it for visualization, analysis, and decision support. However, achieving success implies data provider and consumer trust and buy-in to a sharing culture, agreement and adherence to standards for metadata, vocabularies, and data itself - and/or mapping across varied approaches, and compliance with guidelines (including those on citation and licensing governing content reuse). Agricultural institutions, including CGIAR, are just now beginning to address these issues, and to systematically agree on and adopt adopting standards-based systems and processes, and build cross-walks across differing schemas. Through its Open Access and Open Data initiative funded by the Bill and Melinda Gates Foundation, and via plans for an ambitious Big Data and ICT platform, CGIAR is developing the technical infrastructure that will enable research content to be consistently and seamlessly discovered and interlinked across all 15 Centers. The infrastructure will extend to the agricultural domain in general following the CGIAR pilot phase.

The Allemang and Teegarden paper is therefore timely in highlighting ways in which a data-sharing infrastructure might be built for agriculture by focusing on four areas: stakeholder engagement;

[49] "Progress – Open Data". http://www4.syngenta.com/what-we-do/the-good-growth-plan/progress. n.p., 2016. Web. 15 July 2016.

data sourcing and handling; and sharing and collaboration frameworks.

Stakeholder engagement

The engagement of public and private sector stakeholders within and beyond CGIAR is crucial to the success of efforts to make outputs FAIR and leverage big data capacities for maximizing CGIAR's impact. Efforts center around developing and incentivizing a scientific culture of sharing outputs, and building trust for this by engaging as much as possible with Center divisions and units to advocate and ensure buy-in for OA/OD - from leadership through project management to IT, HR, and knowledge and data management; providing support and capacity to close gaps; implementing clear policy, guidelines and workflows; developing licensing assistance, and more. Engagement with partners who have or can develop analytics, visualization, and decision support tools to add value to existing data is also anticipated to yield valuable dividends.

Data sourcing and handling

Research at CGIAR Centers focuses on different commodities, disciplinary domains, geographies and scales, resulting in very different types of data - some born digital, often characterized by large size and speed of generation, and frequent updates. Data therefore ranges from agronomic trial data collected by field technicians in a variety of ways and formats, through input and output market information and socioeconomic data on adoption and drivers and conditions to enable it, to weather data and high-throughput sequencing and phenotypic information and satellite images. These datasets cannot all be treated in the same manner; the curation and quality control needs differ significantly, for instance - necessitating somewhat customized approaches depending on the data type.

Sharing frameworks

As Allemang and Teegarden point out, there are several data-sharing frameworks to guide data exchange and use. CGIAR strongly encourages adherence to the FAIR principles to render information outputs Findable, Accessible, Interoperable, and Reusable. The benefits of FAIR information resources include transparency, improved efficiency and returns on investment, higher quality data, increased collaboration, economic benefit, and enhanced innovation and impact. CGIAR also aspires with varying degrees of success to the Berners-Lee five-star maturity model to make data available (e.g. via appropriate licensing); structured to be machine readable; in non-proprietary format; referenceable such that it can operate with other linked data; and usable in context through interlinkages across data sets.

Collaboration frameworks

Recognizing the need to democratize agricultural research information and make it accessible to partners in developing countries, CGIAR's aspirations focus on enabling data discovery, integration, and analysis via an online, Semantic Web-enabled infrastructure. This system, built under the auspices of the OA/OD initiative and the Big Data and ICT platform, will harvest content from CGIAR Center repositories to begin with, and include the ability to relatively seamlessly leverage it with existing and new analytical and mapping tools. However, Git repositories on GitHub are also being employed to foster collaboration and share software.

As Allemang and Teegarden note, there is no blueprint for building such an ecosystem in the agriculture domain; however, there are successful models to learn and draw from. Of particular interest are the functionalities demonstrated by the biomedical community via the National Center for Biotechnology Information (NCBI[50]) suite of databases and tools, with attendant innovations for translational medicine and human health. CGIAR efforts to enable similar functionalities to NCBI's are underlain by strong and enduring stakeholder engagement and capacity building, and include many of the areas below that Allemang and Teegarden have proposed.

Incentivizing data sharing by building accountability, culture and trust

Without incentives and a culture that encourages and rewards the sharing of research outputs, any technical attempts to enable OA/OD will meet with limited success, at best. As noted above, the FAIR principles and Berners-Lee's maturity model are at the heart of CGIAR's data sharing framework, and the basis on which accountability, culture, and trust are being built. Among other factors influencing these goals: Sufficient funding needs to be available (e.g., from funding agency support for OA/OD and researchers budgeting and planning from project inception); the benefits to data contributors of sharing their data need to be clear (e.g., increased citations, enhanced collaboration networks, possibility of improved funding etc.), the process of managing data for sharing needs to be easy and well supported (e.g., simple and consistent annotation tools, easy workflows for ensuring data curation, quality, and uploads, guidelines, training on best practices and key tools for data management, support for digital data collection etc.), and the conditions for reuse need to be clear (e.g., citations, licensing etc.) Researchers need to be accountable for making their outputs openly accessible (e.g., through contractual obligation, annual performance evaluation and recognition, donor policies etc.) Likewise, for continued use and growth, data consumers need to have confidence in the open outputs, through rigorous curation, quality assurance, and where appropriate, clear indications of the uncertainty associated with the data. Lastly, consumers need assurance that open content will continue as open "into perpetuity".

Assuring data privacy and security as appropriate

The concern regarding data privacy and security relates to the issues of trust and sustainability addressed here, but is significant enough to merit special attention. Any CGIAR repositories and harvesters of data need to provide assurance of data anonymity when necessary (for sensitive data such as that on water bodies or forest resources that span geopolitical and/or cultural boundaries, conflict-prone zones etc.) Related to these issues is the concern around ethics, particularly when human subjects are the focus of studies (e.g. surveys). CGIAR is paying more attention to the creation of and continued support for Institutional Review Boards (IRBs) or their equivalent at Centers, along with regular assessments of research compliance with guidelines on ethical data collection and handling. Lastly, whether data is closed or open, it needs to be securely held in the face of such threats as hacking and unanticipated loss.

Enabling interoperability through the use of standards

Interoperability is probably the single most important objective of CGIAR's OA/OD initiative. It is critical to enhancing meaning and context through interlinkages between related content types (e.g., data and publications) and across related data types (e.g. an agronomic data set and related socioeconomic data). CGIAR's approach to interoperability focuses on mapping Center repository metadata schemas to a CG Core metadata schema, the use of standard vocabularies (AGROVOC), and strong reliance on ontologies developed across CGIAR (efforts such as the Crop Ontology[51], and in-development Agronomy Ontology, and Agricultural Technology Ontology) and others (ENVO[52], UO[53], PO[54] etc.)

Ensuring sustainability

For a global data ecosystem such as CGIAR's to endure and become a reliable and trusted resource, a business and/or sustainability model needs to be articulated from the start. At CGIAR, this includes not just a funding model, but also a plan for infrastructure hosting, continued maintenance and growth, and governance; data storage, security, and ethical handling; and easy and reliable content discovery and retrieval from anywhere.

CGIAR's ongoing OA/OD initiative, and in-development Big Data and ICT platform are engaging globally with multiple stakeholders to foster a

[50] http://www.ncbi.nlm.nih.gov/

[51] http://www.cropontology.org/

[52] http://bioportal.bioontology.org/ontologies/ENVO?p=classes&conceptid=root

[53] http://bioportal.bioontology.org/ontologies/UO?p=classes&conceptid=root

[54] http://www.plantontology.org/

collaborative, shared approach where possible towards a state-of-the-art infrastructure to enhance discoverability, access, interoperability, and analysis and reuse of agricultural research outputs. Such an effort can only be strengthened from bringing together the best minds and approaches, from sharing scarce resources where possible, and from not reinventing any unnecessary wheels. While the infrastructure may be initially piloted for CGIAR, the overall goal is to transform the agricultural research and agriculture for development sectors into organizations that harness the power of agro-informatics and big data capabilities to fully realize the value of its outputs and donor investments. This can only happen through wide engagement with National Agricultural Research Systems (NARS), International Agricultural Research Centers (IARCs), academia, and other governmental and international public and private sector entities, and approaches that are guided by FAIR principles and the five-star maturity model.

AGRIMETRICS RESPONSE

Silver Oliver

The GODAN paper entitled “A global data ecosystem for agriculture and Food” presents a vision of a data ecosystem for agriculture and goes on to make a series of recommendations outlining the steps needed to realise this vision. Agrimetrics lives within this ecosystem and is working on a number of initiatives that address these challenges. The importance of a cohesive data ecosystem and its ability to mature with time is considered crucial to Agrimetrics and the wider industry as a whole.

About Agrimetrics

Agrimetrics is a not-for-profit company that was set up in the UK in 2015. The organisation’s aim is to support a revolution in the use of big data science in the agri-food industry and contribute to a highly intelligent, productive, efficient, resilient and sustainable system. Agrimetrics’ vision is to use the power of linked data to repair the connections between food producers, processors, retailers and consumers that have sometimes been broken as the food system has transitioned from local to global. Insights obtained through the use of big data and analytical tools will support a detailed and collective understanding of the agri-food system by farmers, food manufacturers, food retailers and consumers. This kind of understanding is key to making the system sustainable and more resilient.

As an illustration of the power of linked data, Agrimetrics has developed a first iteration of its data platform which includes weather, soil and crop data. A range of application programmer interfaces (APIs) have been developed which allow both data and metadata to be queried. By querying the data it is possible to identify locations which have specified growing conditions defined by both weather and soil characteristics. The APIs on the metadata give access to a standardised description of the soil, weather and crop data domain which mean that it is straightforward to link the data in the platform to other data, for example data that is held privately by another organisation. Agrimetrics is now exploring ways in which the platform can be extended.

The recommendations outlined in the GODAN paper relate to the work of Agrimetrics in a number of ways:

Developing new business models

The Agrimetrics business has been established through government funding but must become quickly self-sustaining. Considerable thought has thus been given to the relevant issues raised in this report. The first is the creation of an appropriate business model to support greater sharing of data. There is no doubting that value exists in shared data across the agriculture industry, but the key is to create an appropriate mechanism by which this can be unlocked. Whilst the creation of a market in which money is the counterpart transaction is one option, it is not the only one.

Agrimetrics is developing a system, supported by an application, in which farmers and growers can engage in informal transactions in which data can be shared in a controlled, safe manner in return for insights based on the data that they are prepared to share. Two principles are key to this being an acceptable approach. The first is that original sharer of the data retains control of that data, and the second is that Agrimetrics is trusted to use the data in accordance with a set of clear principles. We believe that this trust can only be established if we are seen to be clearly separate from government and acting in the best interests of small producers and consumers.

Having success stories to share can help build confidence and trust in new business models; GODAN has the potential to be a powerful channel for the sharing of these success stories.

Leading by example

The commercial remit of Agrimetrics means that it must be focused on delivering value to both data providers and data consumers.

Agrimetrics is building an environment in which sensitive data can be kept secure, but interrogated in a way which yields results which are valuable and anonymous. It is also developing data products which can drive innovation in the sector. These include APIs which can be used to integrate into our platform and widgets which can be dropped into applications or dashboards.

Agrimetrics also consumes open and licensed data. GODAN is well placed to be an advocate for making critical datasets of the ecosystem available whether that be open or licensed. Agrimetrics’ experiences indicate that even if stakeholders are willing to pay for data in agriculture, trust may not be sufficient to gain access.

Encouraging data standards

A crucial aim of Agrimetrics is to build its platform to be part of a much bigger global data ecosystem. For this to happen, a setting in which diversity flourishes needs to be created. Agrimetrics fully endorse the view expressed in the report that there is no need for another data standard. Existing data standards need to be made transparent but the real need is for a stitching together of standards. Agrimetrics will contribute to this by creating robust and rich data models of their own and mapping these to external standards as they evolve. Agrimetrics will publish their data models on their platform.

Models arounds vocabularies and reference data are seen as equally important to (if not more than) data models. Providing registries of identifiers is the glue that will hold the ecosystem together. Again this is an area in which Agrimetrics plans to be an active contributor.

Related to the issue of standards for data is that of quality standards. The Agrimetrics approach here is to find ways of communicating uncertainty in the data. One of the features of a data rich world is that the quality of data becomes much more variable. By creating arbitrary red-lines on the basis of perceived quality, the risk is that the initiatives are constrained to a small(er) data world. Whilst a given level of uncertainty may be unacceptable in one application, it might be perfectly alright in another. In fact, as we transition to a world where decisions are informed from multiple directions, these decisions will become inherently risk based.

GODAN can play a key role in facilitating collaboration around existing efforts to develop standards and identifiers. A crucial role is to help organisations working on complementary efforts be aware of each other, as well as providing gap analysis on missing initiatives.

Annotating datasets

Beyond the adoption of standards, practical support is needed around how to apply them. Annotation is an area that can be approached in many different ways and training and support will be needed for groups without the tools or experience to effectively annotate. Training is also needed more widely in data sharing and using linked datasets.

There is potentially a role for GODAN in providing workshops, case studies and training material for those working with data. This should look at the description, publishing and consumption of agricultural data.

Automated data collection

Agrimetrics believes that better data collection is a key part of the ecosystem. This goes further than just automated data collection and will need improved and novel tools for collecting data from people as well. In the FarmASSIST application we collect data directly from farmers at the time of planning a crop providing data about their intended market for (e.g. bread versus animal feed). Data of this type could be complemented with automated collection of data regarding observed yield and weather.

Data hosting

Data hosting is not the primary focus of Agrimetrics, nor do we believe that it should be that of GODAN. Data is hosted where it is hosted. If certain stakeholders need support in contributing to the data ecosystem then Agrimetrics can provide it but it is not thought that a central repository is a valid contribution to an environment in which diversity flourishes.

Conclusion

As reflected in the initial problem statement, Syngenta found that it was publishing data into what is an immature data discipline in the Agriculture space. The principles laid out in the GODAN paper will be essential in developing this ecosystem and Agrimetrics will be directly tackling some of these challenges as discussed. We welcome this report and look forward to playing our role in implementing its recommendations.

ODI RESPONSE: BUILDING THE DATA INFRASTRUCTURE FOR AGRICULTURE AND NUTRITION

Jeni Tennison

Data is infrastructure. It underpins public services, business innovation and civil society. Data such as statistics, maps and real-time sensor readings help us to make decisions, build services and gain insight. Data infrastructure will become more vital as our populations grow and our economies and societies become more reliant on getting value from data to innovate and improve.

This is particularly true in agriculture and nutrition. We face a challenge of feeding 9 billion people by 2050, with global demand for food, feed and fibre doubling in that time period combined with the pressure of climate change. This can only be addressed by enhancing the productivity of fisheries and farms, improving the functioning of markets, and informing the food choices made by consumers.

Data is an essential tool in our arsenal for achieving the changes that are needed in our agricultural ecosystem. It helps us to make better decisions along all stages of the food production chain including financial services, planting, processing, transportation, and getting to market.

Data infrastructure includes datasets; the technology, training and processes that makes them useable; policies and regulation such as those for data sharing and protection; and the organisations and people that build and maintain data. A working data ecosystem enables and is enabled by a strong data infrastructure.

The GODAN/Syngenta paper ‘A global data ecosystem for agriculture and food’ discusses several important aspects of creating this data infrastructure:

- It rightly focuses on the crucial role of trust in building a data ecosystem: in the quality and availability of data, in the institutions that collect, analyse and provide access to it, and in the way in which personal or commercially confidential data is handled and shared.
- It discusses the need to evolve the technologies and standards that are used to exchange data. Importantly this includes how to build data for the web, which enables us to take advantage of network effects.
- It highlights some of the technical challenges that are particularly relevant for data in agriculture and nutrition, such as dealing with high volumes of data generated by satellite imagery and weather predictions and the streams of data that come from sensors in fields and agricultural equipment.
- The report also touches briefly on the need to develop diverse business models which provide incentives and assurances to the entities of different types which collect, analyse and share data. They include commercial services which generate revenue or recover costs, while also providing shared and open data[55]. A global data ecosystem will include organisations operating under a variety of models.

Faced with these issues, it can be hard to know where to start. Creating a strong data infrastructure cannot happen overnight. It takes time to build trust, to develop standards, to embed processes that ensure data quality and availability, to grow new businesses and evolve existing business models.

Below we outline some recommendations about how to approach defining the data infrastructure within a sector. This is relevant for agriculture and nutrition, but could equally apply as a framework for other sectors to learn from.

[55] http://theodi.org/data-spectrum

Think big but start small

It is easy to be overwhelmed by the huge scope of a sector such as agriculture. Starting with a focused real problem — such as how a smallholder can tell when to plant their crops for maximum yields, or how a supermarket can guarantee the contents of a ready meal to its consumers — makes the challenges tangible and addressable. What data is needed to combat this particular problem, who needs it and when? Which stakeholders need to be brought on board and what motivates them? Who currently holds data which could be helpful for solving that problem? Where are the barriers that need to be overcome? What capacities and resources are needed to make sense of the data?

Data often has wide applicability. Having data about the characteristics of a particular crop can help inform not only when it is planted but also what herbicides to use or how to use it in rotation with other crops. The serendipity of unanticipated uses of data is one of the reasons why open data can be so powerful. This can mean that starting with a real world problem feels constraining. Why focus on addressing a single problem when we know that data could be used in many many ways?

The answer is that starting with problems ensures that the approaches that are offered are grounded in the real world. In technical domains such as data, there is often a temptation to try to find a problem that might exploit a particular technology or solution, rather than using the technology that best addresses the problem. Starting with the problem and fully understanding the needs of those that will be using data ensures we aren't tempted to use technologies that don't suit the data we're handling or the people that want to process it.

Starting with the problem also helps direct the limited time and resources that people have into achieving concrete impactful results sooner, which builds greater buy-in over the long term. The fact that the same data and organisations involved in addressing one problem are also relevant for another (and even for problems in other sectors entirely) merely means that tackling one problem makes it easier to tackle the next one. This drives the gradual growth of our data infrastructure and ecosystem.

Consider the whole system

Data operates within a broader system of individuals and organisations with existing processes, collaborations, business models and methods of decision making. We have to understand how to maximise the value of data within such a system, and the potential negative impacts that it may also bring. To get these stakeholders to work together to strengthen data infrastructure we need to find ways to benefit everyone.

At ODI, we use the wheel shown in Figure 3 to illustrate the activities that are needed, in our experience, to tackle collective data challenges within a given sector. We have seen good data initiatives fail because they address only parts of this wheel. A common failure mode is targeting only technical aspects, such as the publication of data and the creation of standards. Both these activities are useful but they are not the only, or even the major, activities that are needed to support open innovation with data.

We have found the following activities are also necessary:

- Creating the tools to support data standards and the techniques to apply them that enhance their adoption.
- Building an evidence base and conducting a user needs analysis to ensure that the problem, and existing approaches, are well understood.
- Building capacity across the sector through training, workshops and the development of assets that help people learn how to use relevant data.
- Engaging the community through media and events so that they know about the activity.
- Developing policies, regulations and legal frameworks that ensure data is handled ethically and transparently.
- Incentivising the use of data in the development of products and services through competitions and startup incubation.

Directing such a change programme to tackle real world problems requires transparent and accountable governance structures, including clear goals and performance metrics, and an overall strategy.

When designing these systemic interventions, we have to take into account that there are large variations between those affected not just in terms of their data literacy and technical capacity, but also in aspects such as education, wealth and power. The people and organisations in the agricultural ecosystem are particularly diverse. They include policy makers, businesses from international agritech firms to startups operating on a shoestring, large and smallholder farmers, CSOs, the media, academic researchers and everyone who eats food.

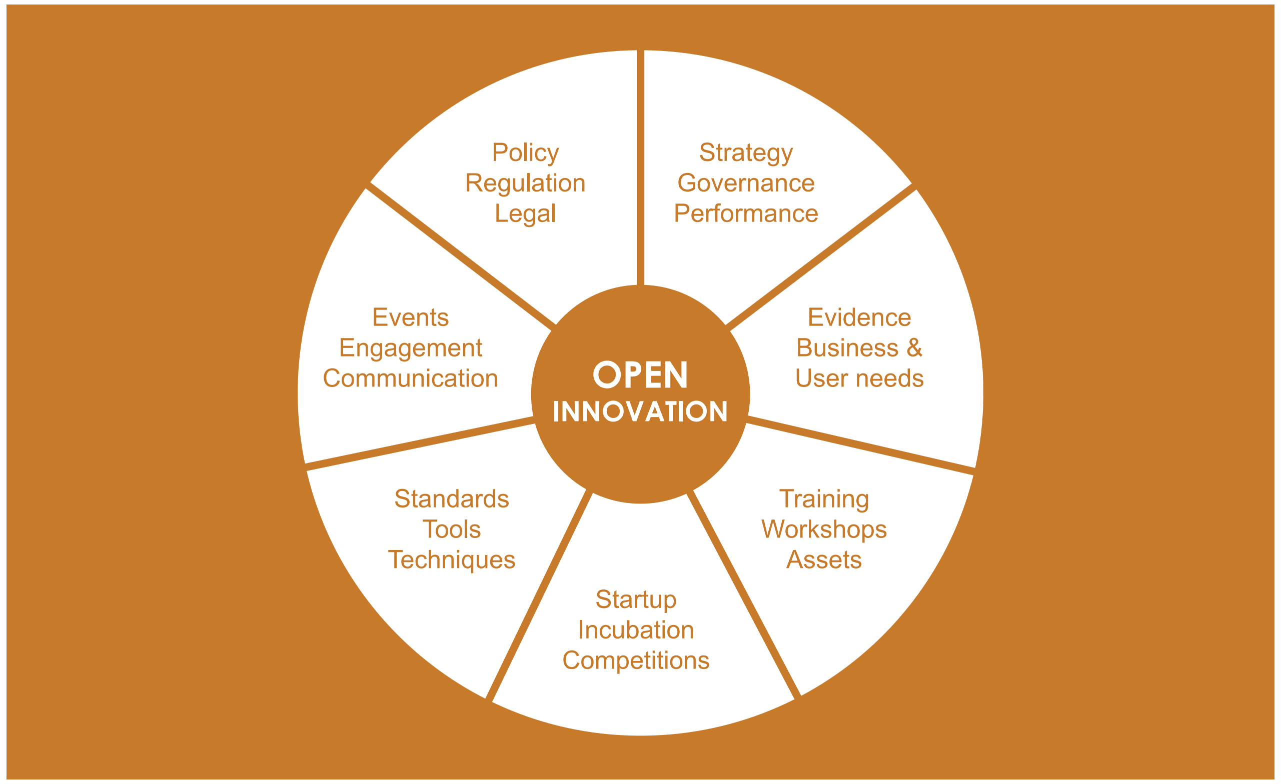


Figure 3: Activities that lead to sector-level change through open innovation

It is easy to design data ecosystems in which the digital divide deepens information asymmetries and makes existing inequalities worse. But data can also help address those inequalities if, for example, it is used by researchers, innovators, charities and individuals to equip people and communities to hold the powerful to account and to inform their decision-making. Building a data infrastructure that brings benefits to everyone requires a wider understanding of the whole system in which that data plays a role.

Harness the network

Increasing the size of a network increases the value it provides. This applies to the collaborative maintenance of data: the more people contribute to maintaining a dataset, the higher quality it achieves and the lower cost it takes for any individual. It applies to trust: the more people reference and use a canonical list, for example of pests or plant diseases, the more valuable that list becomes for linking data together. And it applies to open data infrastructure: the more data is open, the easier it is to get value from the data that we have.

Organisations such as GODAN are essential in harnessing these networks towards real world problems and achieving an initial critical mass which enables the network to grow. For example, they can support network effects in the three areas noted above:

- Collaborative maintenance of common datasets requires convening power to bring together organisations to contribute to that maintenance, defined governance structures that make those organisations comfortable with its neutrality, and technical infrastructure to support the storage and maintenance of the data.
- Canonical data sources and open standards can increase interoperability, but only if people know about them and can easily adopt them. This requires identifying, listing and "blessing" those sources and providing tooling that enables conversion from other codelists or formats into those that are encouraged.
- Open policies and processes enable us to get best value from our data infrastructure but they can be hard for individual organisations to adopt as it is seen as losing a competitive advantage. Both advocacy work and convening power help to ensure that first steps are taken together.

The flipside of network effects is that sometimes they can result in non-optimal solutions. Networks of data creators can concentrate around particular codelists that aren't well maintained or become out of date: everyone continues to use them because they are the thing everyone uses, but increasingly have to work around their limitations. Harnessing networks does not just involve building those networks; it also can include shaping activity towards better solutions.

Design to adapt

We are, globally, at the very beginning of the process of developing our data infrastructure and we don't yet know exactly what works, nor what the future holds. The things that we can guarantee are that new technologies will be developed, new standards will be created, new organisations will grow and existing ones will change their focus. The data infrastructure that we build needs to be able to adapt as we learn, and as the world changes.

We can also be certain that solutions that work in one situation may not work in another. This is a familiar story in agriculture: just as we cannot expect the same fertiliser to bring exactly the same benefits to every crop and soil, we cannot expect the same data to be as valuable to every farmer or policy-maker. We need to adapt appropriate approaches and tools to meet the needs of different local contexts, including being mindful of the political economy (e.g. who has power to make decisions?).

The data infrastructure that we build and the way in which we enable it to develop has to embrace change, adaptation and evolution. There are many aspects to this:

- employing participatory and human-centred design to ensure that approaches are adapted to local contexts
- adopting a scientific approach to measuring the impact of interventions that is oriented around constant research, evaluation and learning (such as those being developed by GODAN Action)
- encouraging innovation, both directly through challenges and incentives, and indirectly through embedding an open culture and makes experimentation easy
- using technical approaches that are built around modular, distributed datasets and services, and using open standards in the protocols and formats that are used to join them

At the same time as embracing change, it is important for data infrastructure to be solid and reliable, so that it can be built on. Building an adaptable data infrastructure is as much about recognising and highlighting those parts that are stable as it is about building in points of extensibility and customisation that enable evolution and growth.

Nurture an open culture

Making a data infrastructure that is as open as possible is beneficial for everyone. Open data, open standards, open source and collaborative models build trust, reduce cost and create more value than other approaches. Being open also increases the number of connections that can be made in a network. As discussed above, data benefits from network effects: it creates more value as more people use and maintain it.

Openness is also important in the process of building a data infrastructure, both in terms of being open with people, so they know what is happening, and being open to people, so that they can contribute to the outcome. Transparency builds trust within the data ecosystem. User engagement enables us to accurately define problems and issues, identify data needs, and articulate the objectives of a programme of change.

Shared discussion papers like this, and thematic working groups focusing on tackling specific problems like data rights, are a good start to being open. But we need more proactive outreach to ensure different voices are brought into the conversation, including the most vulnerable. As we have discussed above, the stakeholders in the agricultural sector are extremely diverse, as are the contexts in which data-based solutions may be deployed. The data infrastructure we build will be similarly varied. Openness with and to the people it is meant to benefit will ensure it can be adapted to their needs.

As the GODAN/Syngenta paper suggests, we can't afford to neglect the issue of culture at the community, organisational, or institutional level. Culture defines how we relate to each other and share information. If there are social, political, economic or historical barriers discouraging people from connecting and sharing data, then no technical solution will help. Transformation initiatives that are based on creating an open culture such as Syngenta's 'Good Growth Plan' or the Open Government Partnership can help to catalyse culture change over time.

Summary of recommendations from ODI

Data infrastructure provides a foundation on which services can be built and decisions made. We need to learn the lessons from the evolution of our road, railway and energy networks in the industrial revolution and ensure our data infrastructure is constructed to maximise benefits to society. Data could contribute more value to our economies and our lives than it currently does, and we need to access that value to meet the global challenges that we face.

Good infrastructure is simply there when we need it but, at the moment, too much of our data infrastructure is unreliable, inaccessible or only available if you can pay for access. Data innovators struggle to get hold of data and to work out how they can best use it, while individuals do not feel that they are in control of their data.

A good data infrastructure is designed to be as open as possible while respecting privacy. It is built for the web, developed based on need, evolves and is dynamic, and encourages open innovation. A good data infrastructure benefits everyone.

We have argued here that building a good data infrastructure requires us to start with real world problems, consider the whole system, harness the network, design to adapt and nurture an open culture. Above all, building a good data infrastructure requires us to start somewhere. Each project in which data is used or shared may seem like a small seed, but it's from those seeds that our data infrastructure will grow.

AGROKNOW RESPONSE: CAN EUROPE LEAD A DATA REVOLUTION IN AGRICULTURE AND FOOD?

Nikos Manouselis

A shared global data space for agriculture and food would propel the industry forward. Information would become more available to all actors seeking to foster innovation. Analytical and decision making tools could incorporate a greater abundance of data sources. A digital economy would arise with online services and applications that use machine readable, interoperable and often publicly shared data. The necessary infrastructure components, including the technology, people, policy and business ones, could seamlessly integrate and work together.

We believe that a transition of agriculture and food into a highly innovative industry powered and disrupted by data and digital services is inevitable, the only question is ‘when?’. In this response to the ‘Global Data Ecosystem for Agriculture and Food’ discussion paper, we explain why we feel that Europe[56] is uniquely positioned to lead and accelerate such a data revolution.

What makes Europe special?

European policies to enable the transition of traditional industries into data-powered ones have been put in place in a decisive manner during the past few years. Three very important policy initiatives are expected to completely change the mindset of people and the environment in which agriculture and food enterprises operate, especially in research. Firstly, the decision to create a Digital Single Market[57] for all EU member states and the adoption of an Open Access policy for all EU-funded research outcomes[58]. The Digital Single Market attempts to tear down regulatory walls and move from a position of fragmented national markets for digital goods and services to a single market across the EU. Secondly, the Open Access policy which has been become a mandatory requirement in the Horizon 2020 (H2020) framework programme. This Open Science policy requests that all research and innovation activities that are funded by public money make available their outputs (both scientific publications and research data[59]) as a public good. Finally, the Public Sector Information PSI) directive[t] which has been encouraging the publication and licensing of digital assets produced by the public sector in the EU in a way that will make possible its re-use for both commercial and noncommercial purposes, in a non-exclusive manner.

These policy actions have been backed up by significant investment in a digital infrastructure that will facilitate storage, management, and discovery of data assets. In the scientific domain, this has led to the establishment of cross-European e-infrastructures such as: OpenAIRE[61] that is connecting open access repositories and cataloguing all scientific outcomes produced within EU projects and beyond; OpenMinTeD[62] that is developing a technological backbone of text and data mining services that may be applied to scientific outcomes; Big Data Europe[63] that deploys big data technologies and analytical engines across industrial sectors; and the overarching European Open Science Cloud[64] (EOSC). In a similar manner, the PSI has been complemented by the deployment of two Open Data portals - one for all data produced by the European Commission and one providing access to data sets from national Open Data portals (both accessible through data. http://data.europa.eu).

Another important factor that has uniquely positioned the EU was the decision to invest into and support the cultivation of a European entrepreneurial spirit that particularly nourishes startups and SMEs[65]. The emergence of the open data culture has also significantly influenced relevant EU initiatives that have been financing small businesses to create digital applications and services. Examples like the Open Data Incubator for Europe (ODINE[66]) and the sixteen FIWARE accelerators[67] (of which, five were focused on the agri-food sector) have played an important role in this, creating an initial environment in which such business efforts may be supported and nurtured. In addition to this, the excellent work that the Open Data Institute (ODI) has been doing in the UK has served as a model for many European countries, showcasing how a data-powered digital economy may be created in a systematic and well thought manner. The ODI approach has been reflected in the way in which many EU initiatives and programmes have been conceived and implemented.

Is this the right time?

What is unique and fascinating today in the European landscape is that this is the first time that the enabling conditions are all there. The advocacy work that the Global Open Data for Agriculture and Nutrition (GODAN) initiative is doing has been instrumental in getting together many of the European stakeholders to align their data infrastructure thinking and agendas. European funding agencies that operate in the global sustainability space have highlighted data infrastructure as a priority in the Belmont Forum’s E-Infrastructure and Data Management coordination action[68]. World-class European scientific organisations, such as the French Agronomic Research Institute (INRA[69]) and the Wageningen University & Research Center (WUR[70]), have decided to join forces, mobilising European scientific communities, and together developing a 10-year roadmap for an open science e-infrastructure in agriculture and food. The work that the Food & Agriculture Organization (FAO) of the United Nations has been doing for decades on topics related to agricultural information management and interoperability has become central to the discussion, providing mature expertise and proven facilitation mechanisms. The Interest Group on Agricultural Data (IGAD[71]) of the Research Data Alliance (RDA) is evolving into a global forum that brings people working on research data management together.

In this fertile context, getting the right people together matters. This is what happened at the ‘Open Harvest’ event in May 2016: a select group of people that represented important stakeholders, including INRA, WUR, FAO, CGIAR, CABI and Syngenta, came together and jointly proposed a set of principles contained in what they termed the Chania Declaration[72]. This is a call to arms for the community, asking for enhanced collaboration between the public and private sector, as well as an alignment of the investments on the infrastructure so that a common data space may be created. The various European Commission Directorates (DGs[73]) that have an interest in the agri-food sector will also need to be be part of this ongoing discussion. They have been invited by this group to sit around the same table and discuss about how existing and future EU funding programmes may be best designed and utilised to achieve this common goal. This is the right time for Europe to lead discussions about the way in which a shared global space may be created, as well as to ensure support for the EU organisations that already are among the leaders of this orchestrated effort.

What challenges do we face?

The ideas and proposed directions covered in this discussion paper reflect to a large extent also the ideas and considerations that European institutions have been discussing and tackling on various fronts. The interventions that the authors propose are relevant and timely. The relevant routes and challenges in the European context to meeting the recommendations of this report are:

Building trust. The EU should embrace and support industry-specific efforts that bring together the public and private sectors. The example of OpenPHACTS shows how a pre-competitive research data sharing environments may be created for the benefit of public and private sectors. In agriculture and food we would initially expect to see such efforts first in a restricted contexts, bringing together the public and private stakeholders in well defined sectors (e.g. the dairy sector) or around specific problem areas (e.g. product / food safety).

Developing standards and linking data. Thanks to the work that FAO has been leading during recent years, there is significant experience and expertise

[56] By Europe, we refer to the European Union (EU) of the 27 member states, as well as well as non-EU European countries that will adopt common principles and be interested to co-invest in such infrastructure. This covers countries such as the United Kingdom (after/if the Brexit negotiations conclude) and Switzerland. Nevertheless, in this article I focus on what I see as a unique opportunity for the EU, particularly because of its commonly adopted legislation, infrastructure investment, and single digital market vision

[57] http://ec.europa.eu/priorities/digital-single-market_en

[58] http://ec.europa.eu/research/openscience/index.cfm?pg=openaccess

[59] http://ec.europa.eu/research/openscience/pdf/openaccess/ord_extension_faqs.pdf#view=fit&pagemode=none

[60] https://ec.europa.eu/digital-single-market/en/european-legislation-reuse-public-sector-information

[61] https://www.openaire.eu

[62] http://openminted.eu

[63] https://www.big-data-europe.eu/

[64] http://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

[65] http://ec.europa.eu/growth/smes/promoting-entrepreneurship/action-plan/

[66] https://opendataincubator.eu/

[67] https://www.fiware.org/accelerators/

in this area. Identifying the gaps in data standards, creating a semantic backbone by enhancing the Global Agricultural Concept Scheme (GACS[74]) developed by FAO (using AGROVOC) with CABI and the USDA National Agricultural Library; and developing good practices, guidelines and tools for publishing and linking data in the context of RDA IGAD - in principle, scaling up successful initiatives and efforts. This vital infrastructural component should be embedded in EU e-infrastructures, such as the European Science Cloud and its agri-food thematic aggregator AGINFRA.

Ensuring sustainability. Large research organisations in Europe often have the resources to set up and manage their own data infrastructures, typically with support from national governments. This is not the case for smaller organisations that produce and publish a large variety of valuable small and heterogeneous data sets. Appropriate models to help these organisations sustain their data resources, services and capabilities are essential. Getting them closer to public e-infrastructure provision is essential, but still the sustainability of services will depend on brokering either government or private sector support.

Providing incentives. The agriculture and food industry is dealing with very important societal challenges and market opportunities. The need to combine various data sources and formats in order to solve them is an inherent challenge that the industry must address in order to develop. Making the potential benefits of data sharing clearer will make it happen more readily only in some cases. Positive reinforcement through accreditation, or payment for associated services will help but cultural barriers to openness in such a rich and diverse environment remain solid in places.

Data publishing principles. This is not going to be a one-off exercise and a one-size-fits-all solution. Each type of problem and project requires different data types and formats to be published. Depending on the availability of standards and the maturity of the specific community, data management and publishing principles will need to be customised and published. Operators of infrastructure services and providers of training and support will need to acknowledge, understand and address this. Those looking at crop trial data, soil data, irrigation data, food recall data, etc. will each need to go through their own journeys.

We believe that Europe is strategically positioned to lead the transformation of the agriculture and food industry; it may coordinate and facilitate the development of the open, interoperable and distributed infrastructure that will help a global ecosystem flourish. This effort should go beyond existing isolated activities and proof-of-concept demonstration exercises. For a data revolution to happen, agriculture and food need a fabric of interoperable and interplaying infrastructure layers that will make data sharing and exchange as natural to us as it is to use the road or rail infrastructure to move from one country to another.

Within GODAN, we have established a Data Ecosystem Working Group where we will kick off and frame this important discussion based, at least to start with, on the important work already done in this paper. Our desire (and expectation) is that this discussion will be broad and inclusive, and involve all relevant strategic stakeholders and forums of the EU (and globally). The active participation of some key European stakeholders in GODAN activities ensures that this involvement will be prioritised and pursued.

[68] http://www.bfe-inf.org

[69] http://inra.fr

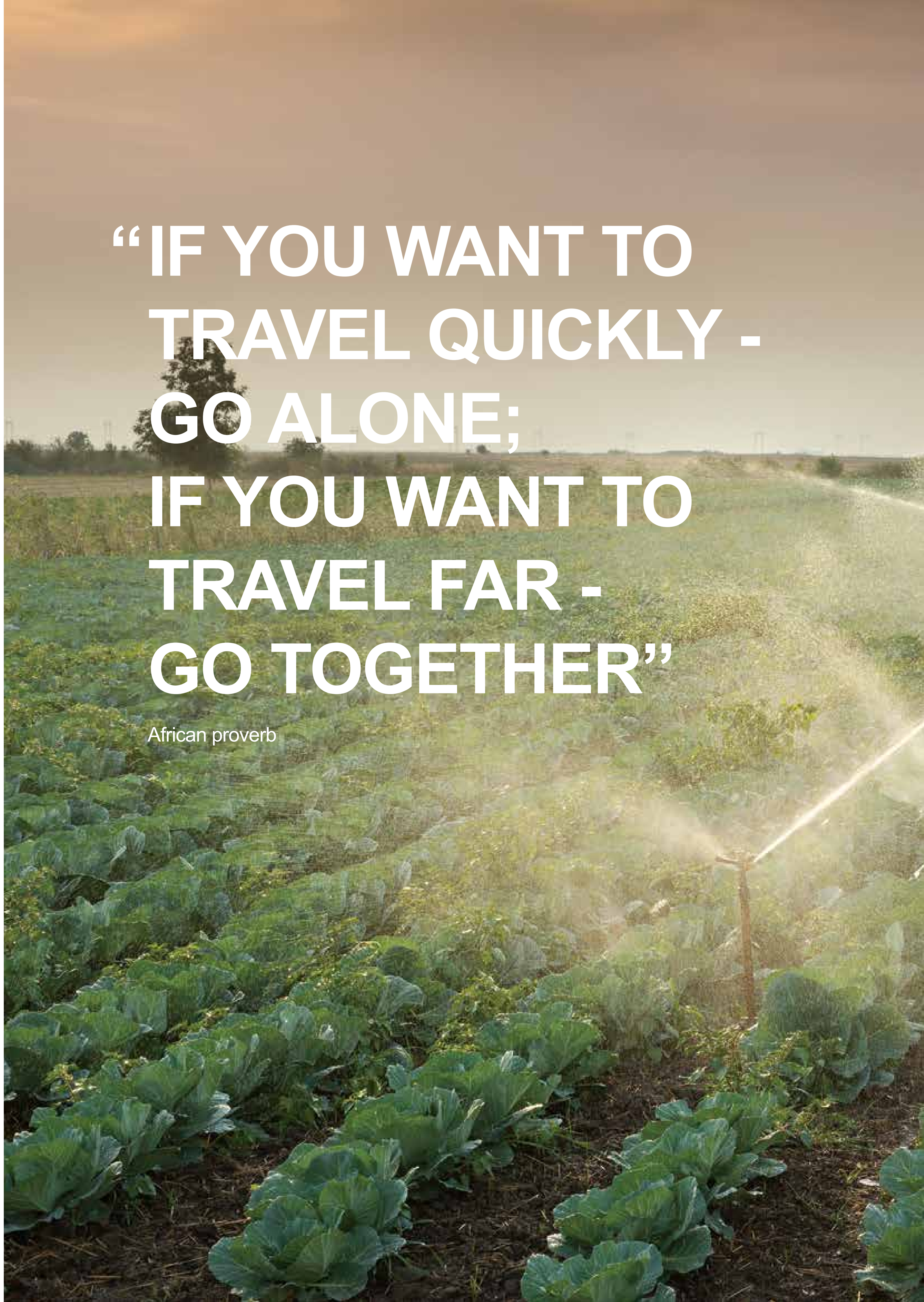
[70] http://www.wageningenur.nl

[71] https://rd-alliance.org/groups/agriculture-data-interest-group-igad.html

[72] http://www.godan.info/news/open-harvest-2016-participants-release-chania-declaration

[73] http://ec.europa.eu/about/ds_en.htm

[74] http://agrisemantics.org/gacs/



“IF YOU WANT TO TRAVEL QUICKLY - GO ALONE; IF YOU WANT TO TRAVEL FAR - GO TOGETHER”

African proverb

Global Open Data for Agriculture
& Nutrition (GODAN)

c/o CABI
Nosworthy Way
Wallingford
Oxfordshire
OX10 8DE
United Kingdom

Tel: +44 (0) 1491 832111
Fax: +44 (0) 1491 833508

Follow us on Twitter:
@godanSec

Thinking of becoming a GODAN partner?
Visit http://www.godan.info/become-a-godan-partner

DOI: 10.1079/CABICOMM-79-12