RESEARCH PAPER

Sustainability in Data and Food

Dean Allemang†

Working Ontologist, LLC., 55 W Livingston Avenue, Columbus, OH 43215, USA

Keywords: Agriculture; Data science; Data sharing; Sustainability; Vocabulary management

Citation: D. Allemang. Sustainability in data and food. Data Intelligence 1(2019), 43-57. doi: 10.1162/dint_a_00005
Received: May 22, 2018; Revised: June 20, 2018; Accepted: July 28, 2018

ABSTRACT

As the world population continues to increase, world food production is not keeping up. This means that to continue to feed the world, we will need to optimize the production and utilization of food around the globe. Optimization of a process on a global scale requires massive data. Agriculture is no exception, but also brings its own unique issues, based on how wide spread agricultural data are, and the wide variety of data that is relevant to optimization of food production and supply. This suggests that we need a global data ecosystem for agriculture and nutrition. Such an ecosystem already exists to some extent, made up of data sets, metadata sets and even search engines that help to locate and utilize data sets. A key concept behind this is sustainability—how do we sustain our data sets, so that we can sustain our production and distribution of food? In order to make this vision a reality, we need to navigate the challenges for sustainable data management on a global scale. Starting from the current state of practice, how do we move forward to a practice in which we make use of global data to have an impact on world hunger? In particular, how do we find, collect and manage the data? How can this be effectively deployed to improve practice in the field? And how can we make sure that these practices are leading to the global goals of improving production, distribution and sustainability of the global food supply? These questions cannot be answered yet, but they are the focus of ongoing and future research to be published in this journal and elsewhere.

1. INTRODUCTION

Projections of world population suggest that we are facing the challenge of feeding 10 billion people by the year 2050. This means that we will need to optimize all aspects of food production, distribution and supply. This problem is not a new one to the agriculture industry; food productivity has been increased dramatically over the past several decades. But many of the methods that have been successful include

† Corresponding author: Dean Allemang (Email: dallemang@workingontologist.com; ORCID: 0000-0002-9947-3427).

methods whose sustainability is in question. The use of genetically modified organisms (GMOs) has come under considerable political attack because of the unanticipated external impact it has on the environment, bringing into question the sustainability of such methods. Conventional methods such as improved chemical fertilizers and pesticides have also dramatically increased productivity, but again at the cost of unsustainable external effects. Poor economic models for these solutions ("they just want to sell us more fertilizer") have also hampered their sustainability.

These approaches need to be applied judiciously and with as much understanding as is possible of their long-term impact on sustainable production. This means that these methods need to be coupled with powerful predictive technology and precision science. We need to tailor treatments of pests to the pests that actually occur, and we need to provide support and fertilizer in the right types and amounts for the crops and conditions under which they are growing. In short, we need to apply big-data predictive analytics to agricultural productivity.

Data can help address virtually all problems in agricultural productivity. For example:

1). **Pest identification**. When a crop is attacked by a pest (e.g., a fungus or mold), data and deep learning technique can help identify the spread and onset of the pest in fields around the world. These methods require large amounts of data for training and can take advantage of image data for fields to provide usable epidemiological data for sustainable crop treatment.
2). **Effective pest management.** What treatments will be effective against an outbreak of a pest? What is the cause of the outbreak? What has worked in the past? What other outbreaks are similar? These questions can be managed with longitudinal data about outbreaks and scientific data about the organisms that cause them.
3). **Planting and harvest planning.** When should we plant a crop, and when will it be ready for harvest? This is not just a matter of timing, but also of climate and weather—how many days of sunshine *vs*. how many days of rain a field has had in a season.
4). **Crop selection**. Which crop should we plant to optimize the productivity of the field? Farmers have been rotating crops for centuries; we now can use big data to make even more productive plans.
5). **Precision farming**. In the developed world, water, fertilizer and pesticides are applied to fields based on gathered data, mostly from satellite images, that analyze the composition of the field.
6). **Crop insurance and finance**. Financial markets set prices based on risk profiles and productivity predictions. More accurate predictions allow for lower insurance premiums and a more stable market.
7). **Research on new pests**. We often observe an outbreak of a pest that has not been seen or studied in detail before. Scientific data on the genomics of invasive organisms can help us deal with pests we have not studied directly before.

All these advances have the potential to dramatically improve productivity. But how do we move from current practice, which includes ancestral knowledge of procedures and decisions which have worked, to applying data-centric approaches that are responsive to changes in the habitat or economic demand? In many cases, we can see improvements on the spot, such as when we use data to understand how to treat

a new pest or understand low crop yields. This sort of solution can be done by a global chemical company (like Bayer or Syngenta), but the application has to happen on the ground, often at farms in the developing world. But more importantly, as we can see from solutions that are already being implemented, the data required for such analysis comes from all over the world, including the developed and developing world, and might rely on historical analysis that requires decades of longitudinal data.

This paints a picture of the data infrastructure requirements for such an undertaking. Data must be collected from a wide variety of sources all over the globe. It has to be shared in a meaningful way, from being able to locate appropriate data sets, to being able to productively interpret the data. The data must continue to be available for decades, even as technology for data management moves on.

This points to a need for sustainable data management. We are not looking at a single corporate data system that will address a particular commercial need, but rather a sustained effort to manage data across the globe. The data cannot be owned by a single organization; like scientific data, this is a resource for research and discovery, and has to be shared across multiple research and operational organizations.

The challenges will only continue to grow. It is relatively easy to increase productivity when current levels are low; optimizing a system that has already been streamlined is much more difficult. We will need to continue collecting and applying data as we move forward. This means that we need to develop a system for creating, gathering and curating data that can continue to evolve.

This paper develops the requirements for such a data ecosystem and outlines the current state. Many of the components already exist and have been in operation for many years; what will it take to make these sustainable through the year 2050?

2. THE UNIQUE NATURE OF AGRICULTURAL DATA

Workers in every industry think that their industry is unique; they have more acronyms to learn, more politics are involved, more competing standards, more personalities who direct research and development in idiosyncratic directions than any other. But for the most part, all industries have these features, and the differences are small. But in the case of agriculture, I believe there is something about the data that really is an outlier in comparison to other industries, and that is how varied the data sources and stakeholders are.

Agriculture is one of the oldest economic activities of civilized humans and has been practised since prehistoric times. The advent of agriculture was one of the great revolutions in how the human race interacts with its environment [1]. In many parts of the world today, agriculture is still practised in ways that are pre-industrial, with human labor using simple tools to work the land. In some parts of the world, farm workers are employed for only part of the year (from planting till harvest), and turn to the large cities, sometimes to become beggars, until the next planting season. At the far end of the spectrum we have precision farming in developed countries, where supercomputers ride on tractors receiving telemetry from satellites that tell them how much of which fertilizer to put on each part of a field. From the perspective of using data to predict and manage crop productivity, the data from each of these extremes are equally

important. If we want to track the progression of an outbreak of a particular fungus, we need to have data from the part-time beggar-farmer as well as the supercomputer on wheels.

In addition to the wide variety of stakeholders, a wide variety of data sources are essential for productivity improvement in agriculture.

1). **Data about agricultural process**. Where did we plant what seed? At what time? In some applications, the data can be so precise that they can be used for future pest control. For example, if we know down to the millimeter where we planted a particular set, we could control for invasive plants by seeing that they are growing a centimeter away. This is like weeding your garden by removing any plant that is not exactly in place. This sort of precise data can be re-purposed during an investigation to know exactly where and when a pest infestation began. The data can be collected during the process (e.g., planting, irrigation and harvest), and typically contain very large data volumes.

2). **Data about crop progress.** How quickly did a crop mature? Has it produced buds yet? Has the fruit begun to ripen? These questions typically can only be answered by observation of the crop itself. This sort of observation can be done manually (sending observers into the field to examine the crop), or by satellite imagery, but more recently it has become possible to gather this sort of data with drones that go into the field and take photos of the crops [2]. These photos can then be analyzed to determine the status of the crop or of an invasive species that has attacked the crop.

3). **Scientific data**. Data about the genome of a crop (which has probably been bred very carefully for certain traits, including robustness against certain known categories of pest) are highly curated and are the result of detailed scientific experiments. These data have large overlap with data about organisms and drugs that are used by pharmaceutical companies for drug discovery and toxicity analysis. This sort of data is often found in peer-reviewed research reports.

   Traditionally in experimental science, a researcher designs and executes an experiment, which includes taking measurements about some process. The experimental setup and measurement policy is documented in the final publication, so that an experienced researcher can replicate the experiment. The data themselves have not been published, for many reasons. The data are typically very large, much larger than the research report or the dissertation itself. Documenting the data in such a way that it can be understood by future researchers is a difficult task, which has to be undertaken in addition to the analysis of the experiment itself.

   More recently, the value of experimental data, beyond the original experiment, has been recognized by a number of academic publishers, and a "data delivery" section of an experimental report is required as a prerequisite to publication.

4). **Historical data**. Data about climate, weather, production, as well as the data about the agricultural process (what crop was planted each year) need to be tracked longitudinally to detect trends and to analyze outcomes of past treatments. This sort of data is already collected for a variety of reasons, but it needs to be integrated with crop data in the context of research of crop failures.

5). **Market data**. Yields and prices play a large role in the distribution and consumption of food, and impact crop planning in future seasons. Insurance and hedges make use of these data to mitigate market risk and to provide a sustainable supply and an economy that can utilize the food produced.

6). **Metadata**. There are at least two main categories of metadata that are relevant to crop protection—schema metadata and vocabulary metadata.

When we consider how we can combine data from different data sets during an investigation of a crop failure, we need to know the structure and semantics of each data set. What are the entities that are being described, and what are the values for those entities that are relevant? In technological terms, this corresponds to the tables and columns of a representation in a relational database, or the elements and tags in an XML message. This is what we refer to as *Schema Metadata*—data about how the data are structured.

When we combine data sets, we will have two or more data sets that refer to the same thing; if this does not happen, then there is not any synergy between the data sets. In the case of pure geographical and temporal data, the common reference is a place at a time; at this time and at that place, one data source informs us that we planted a crop, while another data source tells us that at the same time and at the same place, the temperature had a certain value. The two data sets tell us about what happened at that place and at that time.

But to get real insights about agricultural processes, we need to know a lot more about connection between data sets. We observe a reduction in productivity of a crop in one region, and no loss in another region. Is this the same crop? If multiple companies provide seeds for the same crop, we need to know when we are talking about the same organism. The same goes for pests; if there is an outbreak of wheat rust in this region, is that the same pest that attacked in another region?

This goes for more detailed identities. A single crop can have several phenotypes for the fruit it bears; in classic Mendelian genetics, there were peas that could be smooth or wrinkled. When we get data for a particular harvest, it could tell us that we got 75% of the harvest of smooth peas, and 25% of wrinkled peas. Another harvest might report 50% smooth and 50% rough. Are "rough" and "wrinkled" referring to the same phenotype? We refer to this sort of metadata, such as standard lists of organisms or phenotypes, as "vocabulary data"—it provides a vocabulary for telling us what the data in our data sets mean.

3. DATA-DRIVEN SCIENCE

As it becomes more common to retain and publish the data gathered in an experiment, the basic workflow of experimental science is changing. Instead of starting with a hypothesis, designing an experiment and gathering data, it is now possible to start with data that have already been gathered, then to determine how that can answer relevant questions, and draw scientific conclusions. The process of scientific inquiry is changing from being question-driven to being data-driven. The scientific pursuit of a question is more like a criminal investigation, driven by observations that can support certain conclusions.

This change in scientific activity places some requirements on our data systems. Not only must we gather data, but we have to describe data in such a way that it can be understood and processed by other scientists in other contexts.

As an example of data-driven science in agriculture, Bebber, Holmes and Gurr [3] describe an investigation into an outbreak of coffee rust. The presenting problem is a worldwide outbreak of coffee rust, beyond levels that had been seen before, making a serious economic impact on the market. We need to know the cause of this outbreak so that we can determine what remedy, if any, is applicable.

Three major theories were considered. First, this was a new strain of the organism causing the rust. This line of inquiry was undertaken by the government of Colombia, using ordinary experimental science. Samples of infected crops were gathered, experiments were done on these to compare them to know strains of the organism, and the results were analyzed. The findings indicated that the new infection of coffee rust was caused by the same strain that was familiar from earlier seasons; this was not a new variant of the organism.

The second theory was that climate change was responsible for changes in the habitat that encouraged growth of the rust organism. This research was carried out by the European Bioinformatics Institute, using a wide variety of data, including process-based data (how the crops are planted and tended, moisture, fruit load, etc.), and scientific data about the pathogen itself including for example temperature/response functions. The investigation drew on data that went back to 1987 to do this research. There were issues with finding the data, the medium on which data were stored (often on paper) and interpreting the data for the purposes of the study. The study found that the correlation between measured climate factors and the outbreak of the rust did not have a high enough correlation for this to be the primary cause.

The third theory was that there was some change in management policy in the affected regions. This theory seemed *a priori* implausible, since the change cut across many political regions with different management systems. However, the outbreak happened in the season directly following the global financial crisis of 2008, which had a strong impact on the price of fertilizer around the world. This correlation was high and accounted for most of the change in rust susceptibility that had been observed.

This example highlights many of the issues around data-driven science in the service of agriculture. The investigation is like sleuthing; we do not know at the outset what data will be relevant to the study. There are many independent lines of reasoning, requiring a wide variety of data from around the world and going back for decades. A large part of the research involves finding the data, accessing them, interpreting them and then using them in a new context. It is this sort of example that prompted FORCE 11 to elaborate on the FAIR principles for knowledge sharing. Details of FAIR are given in [4] and [5], 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. In this example, the data in the story are self-selected to have been findable; data would appear in the story if the researchers had not found them. What other relevant data were overlooked during these investigations? There is no way to know.

Accessible means that the data can be obtained by humans and machines. In this example, some of the data were on paper and had to be transcribed. There is a continuum of data accessibility, from paper archives to machine-readable assets. In this story, most of the data were on the "paper" end of this spectrum.

Interoperable means that the data can be managed by a machine and linked to other data sets. This covers parsing standards, self-description of the data and ways to link the data to shared resources. In this story, the data were largely not interoperable; a good deal of the research effort went into linking the data.

Reusable means that the data are described richly enough in terms of shared data (and metadata) resources to make their application apparent. The data in this story were reused, but for the most part, the data sets were selected for their reusability.

If we had a data ecosystem that followed the FAIR principles more completely, the coffee rust research could have been carried out more effectively and efficiently. Furthermore, since we do not know what data were not found by this research, we do not actually know if there might have been different conclusions, or more detailed conclusions, that could have given us more insight into the outbreak. In some research situations, the lack of FAIR data could result in failure of a research project to find an actionable correlation.

4. SINGLE-POINT SOLUTION

One approach to building a FAIR system for management of global agricultural data would be to build a master clearinghouse of all the data and build a custom search index on the data so that any particular data can be easily found. The data could be archived and retrieved on request, making them accessible. The clearinghouse could have a standardized set of metadata (both schema and vocabulary), and all data that enter the clearinghouse would go through a sort of extract-transform-load (ETL) process that would align all data sets with this metadata, making all of the data interoperable. Having all the data aligned under one roof would make data reuse a routine activity.

This sort of single-point solution is not inspired by the success of the World Wide Web, and in fact, this approach is basically antithetical to the principles that allowed the Web to become the largest distributed system of data in history. It should come as no surprise that this approach is fraught with issues—most of which become apparent when we consider the sustainability of the solution.

The most obvious issue is cost. A system of this sort will be expensive to develop (developing all of the ETL protocols, data representation and search software is just the start). There are also repeating costs; the amount of data is massive, requiring the maintenance of hardware and network connectivity to maintain the data. How could this be sustainably funded?

This could be done as a government effort, but the provision and use of the data is necessarily global in scope. What happens when one government decides that another one is hostile? Do we start refusing data access for entities that are not among our allies? This is not how healthy, productive science works. Another problem for government is change of policy when an administration or regime changes. A government that is sympathetic to the collection and dissemination of data could easily be replaced by one that is not.

This could be done as a for-profit company. The undertaking is massive, but one could perhaps imagine a company of the size and capability of Google undertaking the effort. But is it possible to find a sustainable

business model that will allow this to continue? Ubiquitous as Google seems, there could still be a day after the downfall of this giant corporation. Who will continue to curate and support the data after the corporation decides that it is no longer cost-effective?

Another approach would be a not-for-profit but also non-government organization. This would rely on donations and grant support from philanthropic sources. This sort of support is highly competed for, and a data clearinghouse could one day fall behind another effort in priority for limited funds.

Universities provide another source of not-for-profit support, and they have the advantage of durability—many universities are hundreds of years old and show no sign of disappearing any time soon. But university support for efforts of this sort come from funded projects, and when they finish, students and faculty both move on to other efforts. While universities themselves are sustainable, university projects are notorious in how short-lived they can be.

Given the undisputed success of the World Wide Web as a decentralized distributed system, and given all of the issues with a centralized approach to sustainable data management, why would anyone think that the latter would be appropriate? It is not so much an issue about what researchers and developers think is an appropriate approach that limits the way we approach this problem, but rather the ways in which projects are organized and funded. If we see a problem, the mechanism we have (either in a university setting or in a corporate setting, using a not-for-profit or a for-profit model) is to charter a project to resolve it. "Let's build a data sharing platform for agriculture", "Let's build an index to all the ontologies we know about." These efforts are useful and appropriate, but it is difficult to fund or even specify such a project, not as a point solution, but as part of a sustainable ecosystem. The Web was unique in its inception that it was based on technologies that allowed for its adaptation and growth, as opposed to technologies that were focused on the performance of a single system.

How can we run a project, or build a system, with the long haul in mind? It has been done before—this is exactly how the Web standards were built. This suggests that we look to the Web standards for data sharing as a means for building a sustainable data infrastructure for food and agriculture. In particular, the W3C standards Resource Description Framework (RDF) [6] and Simple Knowledge Organization System (SKOS) [7] are particularly suited for sustainable data management:

1). It should be possible for data to be published in a standard form. Both XML and RDF satisfy this standard. In principle, it would be possible to store relational databases in a vendor-neutral form and exchange data in this way. But in practice, the relational database industry has not settled on a method for this.
2). It should be possible to link data sets. In particular, published data should have a global resolution of identity. This is where RDF surpasses simple XML publication, since it has a standard way to take advantage of the identity infrastructure of the Web.
3). It should be possible to share vocabulary to describe the meaning of data. SKOS uses RDF as its basic representation, gaining all of the advantages listed above, but adds to that a standard way to describe controlled vocabularies, making it possible to describe shared meaning.

These standards provide an infrastructure in which data systems can operate, so that data that are provided, curated, vetted or published in any way can be exchanged with other systems, in the present or in the future.

From a sustainability perspective, the data have to transcend any organization that is chartered to obtain or maintain them. This dynamic is exactly how the Web works; even as individual pages on the Web come and go, the Web itself continues. The challenge now progresses from one of finding an organization that will support the data, to one of making an ecosystem that will allow the data to remain and be retained in the face of organizations coming and going.

5. A GLOBAL DATA ECOSYSTEM

In [8], Allemang and Teegarden describe a global data ecosystem for agriculture and food, in which the data for agriculture are not kept in a single place, but rather are distributed across the Web, and supported by a variety of competing and collaborating entities, including governments, non-governmental organizations (NGOs), universities and private enterprises. Instead of viewing different projects or companies as competing in the space of managing the global data infrastructure, it sees them as collaborating as the current stewards of a continuously evolving data ecosystem.

Technology plays a large role in the success of such an ecosystem, and the success of the Web as a source of information and data has normalized the idea that information can and should be distributed, with a variety of people and agencies responsible for the maintenance of that information. But creating and maintaining a global data ecosystem is more than just putting all the data somewhere on the Web. There are challenges to maintaining sustainable FAIR data that go beyond simple Web access and search engines.

Findable. Putting data on the Web will submit them to indexing by major search engines, making them findable to a small degree. But search engines rely to a large extent on the content and interconnection of the things they index. Try searching for the data in an experiment where the prevalence of a particular phenotype in a harvest was measured, when we controlled for the distribution of the genotype of the seeds. In order to search for this sort of data set, we would need detailed metadata about the structure of the experiment, including what was measured *vs.* what was controlled. Putting data on the Web is a step forward from having them on paper, but it does not solve the findability problem. Progress toward this kind of indexing of scientific data has been reported in [9].

Finding relevant data is such a key aspect of any data ecosystem that it should come as no surprise that the state of the art includes a number of indices for data and metadata sources for agriculture. These indices each have their own purpose, and we expect to see such indices continue to evolve.

Accessible. Accessibility of data includes the current accessibility, but more importantly, the sustained accessibility of the data. We can host the data today, but we have to be able to reach them in the future as well. Since many of the data sets are very large, it is not always practical to keep all data accessible at all times. At the moment, many projects and companies provide access to a large number of data sets, but

these cannot be taken for granted; any data set could be retired at any time. The Open Data Institute provides certifications [10] for agencies who want to make claims about the long-term accessibility of their data; any agent who wants to make a promise about continued accessibility of their data can claim a certification level (bronze, silver, gold or platinum), make clear their intentions for their data. While this certification provides a way for organizations to promise ongoing data accessibility, in the current state of the art, few organizations do.

Interoperable. Everyone who publishes a data set hopes that it can provide value to a future researcher and contribute to data-driven science. But interoperability is key to getting added value from a reused data set. This means that data sets have to be published in such a way that various parts of the data can be referenced and cross-referenced to other data sets. The mindset of data provision has to change from "need to know" to "responsibility to provide"—when you publish data, you have to make it possible to refer to them from other data sets. This is both a technical and a research challenge; what technology can be used to annotate a date set so that it can interoperate, and how do you go about doing that annotation?

Reusable. The grand prize for shared data is of course reusability—can you actually use the data in a new context that gets value from them? We see examples of this throughout agricultural innovation, and we saw many such examples in the coffee rust example above. But to a large extent, data reuse involves heavy-duty human intervention. In the case of the coffee rust, a human researcher reused the data in the new situation, as part of an investigation into the source of the rust. Optimizing this process remains a challenge, and few systems today automate this in a meaningful way.

6. STATE OF THE ART

Despite the issues identified in [3] with using global data to support research of coffee rust, the state of the art of the global data ecosystem in agriculture today is better than can be expected. Many pieces of the ecosystem already exist and are interoperating in a productive way.

There are already a number of small enterprises that have made business models from curating, managing and leveraging agricultural data. Agroknow has made a business model for a decade helping organizations and enterprises around the world produce more value from all types of agriculture and food data. Kisanhub provides a data-driven decision support system for farmers. FactBio.com provides tools for the curation and management of scientific data in general, including data for biology and agriculture. In a more general setting, data world has built a platform for crowdsourcing of data sets, including for agriculture. These companies represent only a drop in the bucket of private business models centered around agricultural data; by the time this paper goes to press, any such list will be out of date.

Rather than try to choose among these companies to find a champion, a data ecosystem needs to embrace them all; the data have to outlive the utility of any particular business model.

University projects have produced and curated a number of data indices in agriculture. At the moment, Agrimetrics, hosted at the University of Reading (Rothamsted) provides a federated data ecosystem in

support of agriculture and nutrition. Projects of this sort have a limited period of performance, after which funding will no longer be available and the support will end. But new projects will take their place. Agrimetrics uses standards and data sets developed from earlier projects, and subsequent projects will build on the standards of Agrimetrics. It should be seen as a way station in the continuing evolution of the data ecosystem.

Non-governmental organizations have been involved in the management of agricultural metadata for decades. The United Nations Food and Agriculture Organization (FAO) publishes Agrovoc, a vocabulary for describing agricultural literature and data. The first version of Agrovoc was available in the early 90's, and updates are still being published in 2018. The FAO has joined forces with other vocabulary projects to develop a more comprehensive vocabulary to manage data sets as a global distributed resource. The CIARD Ring[①] catalogs over 3,000 public data sets in agriculture. The value of these resources continues to be recognized and supported by private and public funds, and their success has spawned a variety of metadata management systems. Over 100 ontologies are shared in the AgroPortal described in [11].

The CGIAR[②] advances scientific research and innovation to help poor people around the world share in economic growth. They manage data from all over the developing world, and make them available for innovation and research.

All of these entities and others recognize the need for data to support innovation in our agricultural world. They continue to create, classify and curate data from all over the world.

Small businesses are not the only ones who have found a profit motive in agricultural data. Not surprisingly, agriculture business giants like Syngenta and Monsanto (now Bayer) have found that data are key to the continued innovation of their business. Among other activities, they have founded the AgGateway consortium, which supports the transformation of the world to digital agriculture—agriculture that takes advantage of information technology to optimize production and minimize loss. Syngenta has taken data accessibility and sustainability seriously, and has published at the ODI "Silver" level (see [9]) data about their progress toward sustainable agriculture.

7. TECHNOLOGY

To a large extent, the problems with creating and maintaining a global data ecosystem are organizational and social; how can we find funding models that will sustainably support the curation and availability of data sets long enough for them to be reused (after all, it is not uncommon for a 30-year-old data set to be useful). But there are also technical challenges. We can look at the state of the art to see what technologies have been used to address these challenges so far.

The findability of data can and has been addressed using standard Web technology; putting a data set on the Web with some keywords will allow it to be indexed by the same search engines we use in our

① http://ring.ciard.net/datasets

② www.cgiar.org

everyday interaction with the Web. This is certainly done and is probably the most common way that a search for data sets of any sort begins. There is no reason not to do this, or to underestimate its effectiveness.

But as we have seen, differentiating between data sets cannot be done easily with this method. It is difficult to tell the difference between similar experimental or measurement setups, where different variables were controlled and measured. For this reason, in addition to the publication of data sets, we see the publication of portals, like the CIARD ring and the Ontology Lookup Service of the European Bioinformatics Institute (EBI), which provide search tools specifically for data sets and metadata (ontologies).

What is it about these search engines that makes them more effective than the general search engines (Google, Yahoo!, Bing, etc.) that we use throughout the Web? There are a number of particular features that make them stand out.

Since they are working over a limited domain (food and agriculture), they can use a limited vocabulary in a more specific way. When we use a word in a general search engine, the general use of the word takes precedence over the meaning that is specific to agriculture. An agriculture-specific search engine can use a controlled vocabulary that is specific to agriculture.

The AGROVOC vocabulary is exactly such a thing, and it has been used to index agricultural literature for decades. This provides a strong backdrop for specific search engines. It also provides a way for data sets to standardize on how they refer to common terms in agriculture.

Vocabularies like AGROVOC play a special role in the data ecosystem. They provide touch points where multiple data sets can refer to the same thing, and anyone who reuses those data sets know, without reading through documentation or contacting the data set authors, that they refer to the same thing. They are key shared resources that allow other resources to interoperate.

This means that vocabularies like AGROVOC have to have some very special properties; not only is it necessary to accurately find the vocabulary itself, but it has to be possible to reference each term in the vocabulary in a global, unambiguous way. The FAO team that develops AGROVOC achieves these goals using SKOS. SKOS is based on RDF, which allows each term to be referenced as a Web resource. On top of that, SKOS provides standard bibliographic metadata for each of these terms. This allows AGROVOC and other vocabularies to serve as (meta-)data sets in their own right, providing cross-references between other data sets. Search engines can take advantage of these vocabularies to run focused searches.

The centrality of the Ontology Lookup Service (OLS) [12] indicates another technology that is widely used in the sharing of data for agriculture, that is, technology for the representation and management of ontologies. The OLS currently indexes over 200 ontologies, most of which are published using the W3C Web Ontology Language (OWL) [13]. OWL builds on RDF, so it shares with RDF the ability to reference parts of an ontology uniquely and globally. To this it adds the ability to describe the relationship between terms in a logical manner. This makes it possible to describe the relationships between genomes and phenotypes, between yields and prices, how pests interact with the crops they attack, and how various

treatments work on those pests. A great deal of scientific data has been represented in OWL and is indexed using the OLS and other such tools.

Together these tools currently provide a wide variety of information that begins to approximate the FAIR principles; we can use vocabularies like AGROVOC to find data sets, and to describe how they interoperate. We can use RDF and other technologies to represent data sets, so that they can be accessed in a uniform way. We can use OWL to describe the details of the relationships within complex data sets, enabling their reuse.

8. CONCLUSION

Sustainability of our food supply depends on productivity optimizations, which in turn require a sustainable supply of data. Creating and managing the infrastructure for this data source is not a simple matter; this is not a single project that some government or corporation can manage on its own. To be sustainable, a data ecosystem has to include pieces that can evolve and be replaced, while maintaining the integrity of the data content.

This ecosystem exists to some extent today, as small enterprises have found profit models for data curation and sharing, and NGOs have been established with charters for maintaining and utilizing data, especially for the improvement of productivity in the developing world. National governments and the United Nations have played and continue to play a role in managing data and metadata for agriculture.

Just like the Web in general, we cannot expect there to be a single place where the data infrastructure for agriculture data will reside. But simply putting the data somewhere on the Web is not sufficient to support the sustainable reuse of data in a time frame that will keep up with the sustainability needs of the food supply.

The current state of affairs is not sufficient to provide the data sustainability that will be needed to continue to grow the food supply, but the groundwork is there. Through a combination of business models and other forms of support, we already have a broad data infrastructure for agriculture and nutrition. Using both conventional and advanced Web technologies, we have already begun to build a network of interoperable data and metadata resources, which interoperate to provide innovative solutions to food production optimization problems. This provides a starting point and groundwork for the development of the ecosystem yet to come.

REFERENCES

[1] Y.N. Harari. Sapiens: A brief history of humankind. New York: Harper, 2015. isbn: 9780062316097.

[2] S. Evans. ACFR: Robots set to transform the automotive and agricultural industries. Available at: http://mar-ketclarity.com.au/acfr-robots-set-to-transform-the-automotive-and-agricultural-industries/.

[3] D.P. Bebber, T. Holmes, & S.J. Gurr. The global spread of crop pests and pathogens. Global Ecology and Biogeography 23(12)(2014), 1398–1407. doi: 10.1111/geb.12214.

[4] P. Kersey. Plant omics data: Emerging standards for data discovery. In: EBI Industry Workshop: Ontologies in Agriculture, Food and Nutrition, 2017.
[5] FORCE11. 2014. FAIR guiding principles. Available at: https://www.force11.org/node/6062.
[6] The World Wide Web Consortium (W3C). Resource Description Framework (RDF). Available at: https://www.w3.org/RDF/.
[7] The World Wide Web Consortium (W3C). Simple Knowledge Organization System (SKOS). Available at: https://www.w3.org/2004/02/skos/.
[8] D. Allemang, & B. Teegarden. A global data ecosystem for agriculture and food. Available at: https://www.godan.info/sites/default/files/documents/Godan_Global_Data_Ecosystem_Publication_lowres.pdf.
[9] H. Rijgersberg. Semantic support for quantitative research. PhD dissertation, Vrije Universiteit van Amsterdam, 2013. Available at: http://dare.ubvu.vu.nl/handle/1871/40428.
[10] Open Data Institute. ODI/Open Data Certificate. What you need. Available at: https://certificates.theodi.org/en/about/whatyouneed.
[11] RDA Agrisemantics Working Group. September 2017. Landscaping the use of semantics to enhance the interoperability of agricultural data. Available at: https://www.rd-alliance.org/system/files/documents/Deliverable1%20-%20Landscaping.pdf.
[12] The European Bioinformatics Institute (EMBL-EBI). Ontology Lookup Service. Available at: https://www.ebi.ac.uk/ols/.
[13] The World Wide Web Consortium (W3C). Web Ontology Language (OWL). Available at: https://www.w3.org/OWL/.

AUTHOR BIOGRAPHY

Dean Allemang, best known as co-author of *Semantic Web for the Working Ontologist*, is a leading expert on semantic data integration in the enterprise. Over decades of working with knowledge-based semantic data systems, he has contributed to a variety of industries including finance, media, health care, government, drug discovery and agriculture. He is expert in the full stack of W3C Semantic Web technologies, including RDF, RDFS, OWL, SPARQL, SKOS, and SHACL. In recent years, he has brought his experience working with semantic systems in finance to bear for the EDM Council's Financial Industry Business Ontology (FIBO) standard. He has also deployed semantic systems in a major media firm. His research interests lie in how to use linked data systems to support feeding a growing global population through the next century. With a Master's Degree in Mathematics from the University of Cambridge and PhD in Computer Science from the Ohio State University, he brings a formal approach to semantic modeling. He has developed and delivered courses for the Semantic Web in a variety of corporate settings, and has trained over a thousand practitioners in Semantic Web technology.