Food and Agriculture Organization of the United Nations

E-AGRICULTURE IN ACTION: BIG DATA FOR AGRICULTURE

E-AGRICULTURE IN ACTION:
BIG DATA FOR AGRICULTURE

Edited by
Gerard Sylvester

Published by
Food and Agriculture Organization of the United Nations
and
the International Telecommunication Union

Bangkok, 2019

Required citation
FAO and ITU. 2019. *E-agriculture in Action: Big Data for Agriculture.* Bangkok.

ISBN 978-92-5-131659-7

FAO and ITU information products are available on the FAO website (www.fao.org/publications) and ITU website (http://www.itu.int/en/publications) and can be purchased through publications-sales@fao.org and sales@itu.int

Cover photo:
Cover: *Abstract data vector* – © Adobe Stock/Tartila. Farmer – Shutterstock/Kosin Sukhum.
Back cover: © Adobe Stock/Lamyai

Contents

Preface

The Food and Agriculture Organization of the United Nations (FAO) and the International Telecommunication Union (ITU) continue to work together to promote the use of sustainable information and communication technologies (ICTs) for agriculture.

FAO and ITU follow a three-pronged approach to assisting member countries in identifying, developing and implementing sustainable ICTs for agriculture. First, through the development of the national e-agriculture strategy and related linkages creation; second, is through the support to solution implementation together with national partners and third, in promoting knowledge sharing through knowledge products such as the **E-agriculture in Action** series of publication and the bi-yearly E-agriculture Solutions Forum.

This publication on big data for agriculture is the fourth in the E-agriculture in Action series of publication. With the growth of technology including the impending introduction of 5G networks, which will support a huge sensor network infrastructure, data driven agriculture and the challenges of extracting meaningful insights from various data streams to influence policy decision and/or provide actionable advisories for agriculture stakeholders are gaining prominence. This publication tries to shine some light on how various organizations address these challenges.

The articles in this publication are written by the respective authors and are entirely their own views. We have tried to maintain the original narrative style of each contributor. Neither FAO, ITU nor the CGIAR Platform for Big Data in Agriculture promotes or endorses any of the statements, comments and products mentioned in the articles. Thus, this is an effort to share knowledge on the use of successful ICTs for agriculture initiatives and we expect that this compilation of case studies will be read in that spirit.

Acknowledgements

This publication is the fourth in the series E-agriculture in Action. This is no mean task and this is all due to the wonderful support of the authors and their organizations and their valuable contributions.

FAO and ITU are very grateful to CGIAR Platform for Big Data in Agriculture for their partnership in bringing out this publication. The importance of sharing knowledge on the use of emerging technologies cannot be overstated.

The editor is grateful for the support provided by Kundhavi Kadiresan, Assistant Director General and Regional Representative, FAO Regional Office for Asia and the Pacific and Ioane Koroivuki, Regional Director, ITU Regional Office for Asia and the Pacific for their guidance and support.

FAO and ITU extend their special thanks to the authors and their organizations – Dharani Burra (CGIAR Platform for Big Data in Agriculture), Sanjay Srivastava (ESCAP), Zoe Maddison (Olam), Nina Getachew (FHI360) and Kathryn Alexander (Development Gateway), Julie Cheng (Harvesting), Bessie Schwarz and Anindita Chakraborty (Cloud to Street) and Hiroshi Uehara and Atsushi Shinjo (WAGRI).

We thank Iljas Baker for the language editing.

© Tom Fisk from Pexels

Article

1

Data driven agriculture: the big data phenomenon

According to Forbes, we generate almost 2.5 quintillion bytes of data every day (Marr, 2018). With more than half of the world online, as of 2019,[1] the amount of data generated is colossal (Figure 1). The increase in the use of Internet of things (IoT) will only add to this data deluge. The Economist magazine rightly called data the world's most valuable resource (The Economist, 2017), and others have called data the new oil (Reid, 2017). From various incidents in the past, the world has realized the power that an organization or individual wields from being able to access valuable data and then monetizing it or, worse, misusing it for personal and political gain.

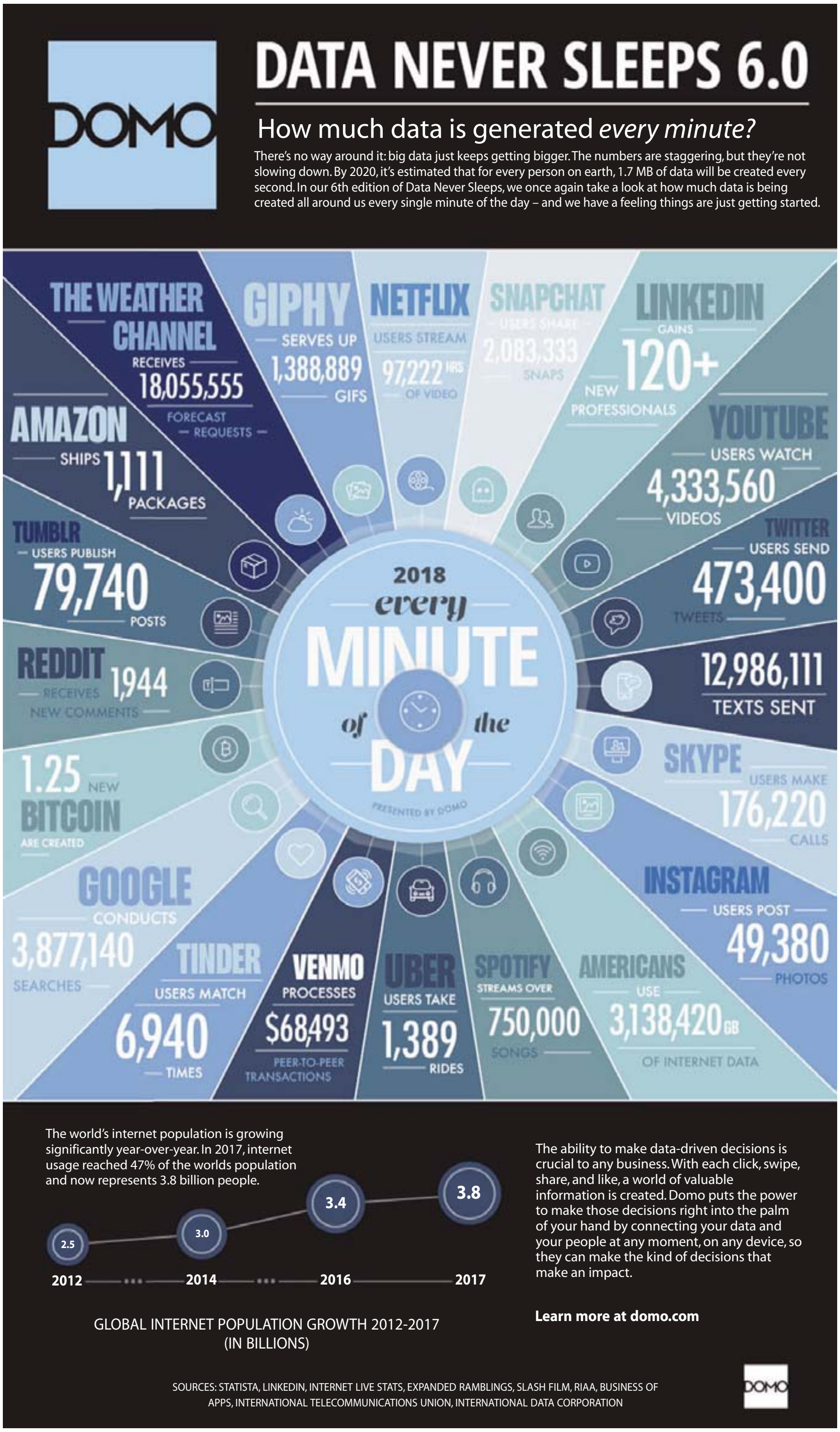
> For sustainable development and humanitarian practitioners, big data and new technologies hold great potential to help measure the effectiveness of projects and programmes, and proactively adjust their implementation based on the realities on the ground.
>
> **UN Global Pulse**

Big data is complex and this brings us to the next big question – how to make sense of these huge amounts of data? How do we analyze data patterns to extract actionable intelligence? What data do we store, what do we ignore? It comes as no surprise that data scientists are among the most sought after professionals (Holak, 2019). The demand is driven by the need to harness valuable insights from data patterns (usually assisted by machine learning) to influence policy decisions or understand behavioral patterns.

How big is big data?

One of the most recognized definition is the one coined in 2012 by the Gartner Corporation that defines big data as "high-volume, high-velocity, and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation".[2] Big data thus requires not only access to large data assets, but also the competence and infrastructure to process them in a timely manner, and the capacity to realize the valuable insights extracted from it (Ali *et al*., 2016).

1 https://www.internetworldstats.com/stats.htm
2 https://www.gartner.com/it-glossary/big-data/



Source: https://www.domo.com/learn/data-never-sleeps-5

Figure 1: Data never sleeps

The four V's that characterize the data assets are:

Volume: the sheer volume of data available nowadays is enormous and growing. This includes data generated by people as well as the billions of sensors all over the world that are creating data every second and communicating with servers over the Internet, creating what is called the Internet of things.

Velocity: the speed at which the data are created, stored, processed, analyzed, visualized and acted upon has increased to up to real time. Big data usually involves collating data generated at various speeds and moments and accommodating bursts of activity.

Variety: the different kinds of data assets keep increasing. These include structured and unstructured data from databases, devices and sensors, logs, social media, websites and posts, images, email communications, and audio and video streams (such as radio and television), among others.

Veracity: the quality (accuracy) of the data and not just the quantity is important and plays a major role in extracting intelligence from big data. The trustworthiness of the data source and processes to "clean up" data by removing abnormalities and inconsistencies are important to improve the accuracy of data.

Many organizations and individuals also extend the definition to include:

Volatility: is the rate of change and lifetime of the data. This also determines the storage period of the data.

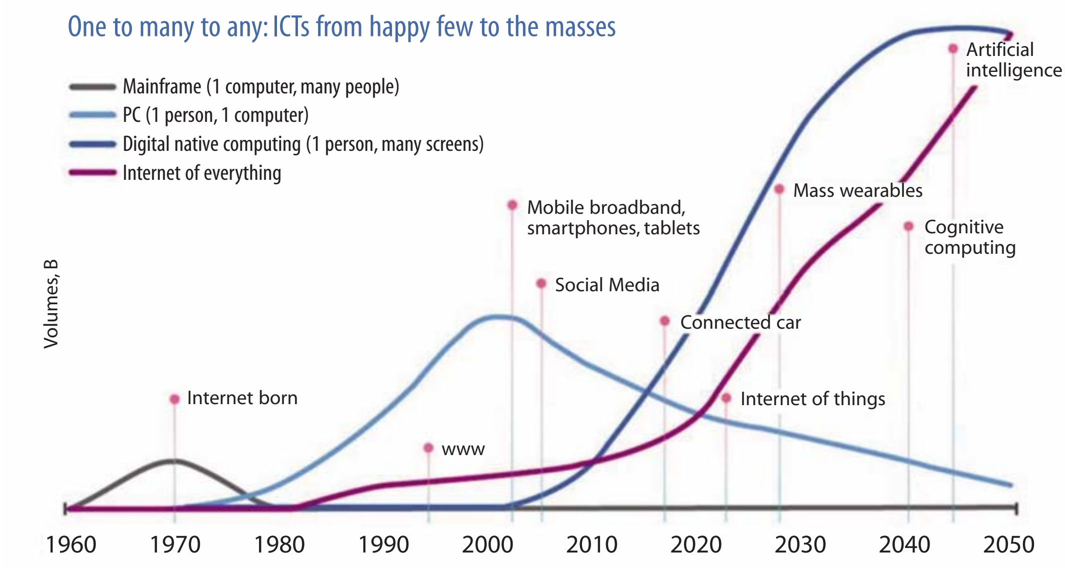
Validity: As with veracity, validity ensures that the stream of data is correct and accurate for the intended use at the desired time.

Visualization: A picture speaks a thousand words, thus the desirability of using graphs and infographics to convey information from complex data patterns.

Value: Getting value from the data is the key objective of big data analysis. How do you transform mountains of data to actionable insights to support your decision-making?

The context of the content is very important. Combining data from multiple data streams and devices should be based on a definitive understanding of the context of the data. The introduction of 5G networks brings in a huge potential to support massive sensor networks, enhanced broadband access together with ultra-high reliable low latency mobile networks. These developments are bound to generate huge amounts of data starting from smart cities, smart transport networks to agriculture and personal wearables, creating the Internet of things (Figure 2). Investment in developing efficient processes and infrastructure is needed to support these technologies and ensure the efficacy of the data generated.

Data harmonization is crucial as data can be structured and unstructured in huge volumes and may be collected from heterogeneous sources. After data harmonization, the time and energy required to run analytics and extract insights from big data greatly reduces. Standardization of big data technologies would also facilitate ease of interoperability, and



Source: International Telecommunication Union

Figure 2: Growth of connected networks

would facilitate generating newer insights by combining or correlating two different data sets. The International Telecommunication Union is working with partners to standardize big data related activities.[3]

With big data comes big challenges

Data privacy

Data privacy, related security and other aspects of handling data are of paramount importance and the regulations related to data privacy in each geographical area have to be taken into consideration. The European Union General Data Protection Regulation (GDPR)[4] outlines regulations and the European Union law on data protection and privacy for all individuals within the European Union and the European Economic Area. The Association of Southeast Asian Nations (ASEAN) framework for personal data protection provides some guidelines on data protection for ASEAN member countries.[5] A report[6] from Deloitte highlights the significance for business of data and privacy protection in ASEAN countries. In particular, it mentions that globalization and digitalization have become a double-edged sword as businesses attempt to comply with data protection regulations in a borderless Internet world. Not long ago, the places where data were generated processed, analyzed, stored, and used were not necessarily on the same continent. This is fast changing now with many countries

[3] https://www.itu.int/en/ITU-T/techwatch/Pages/big-data-standards.aspx
[4] https://gdpr-info.eu/
[5] https://asean.org/storage/2012/05/10-ASEAN-Framework-on-PDP.pdf
[6] https://www2.deloitte.com/content/dam/Deloitte/sg/Documents/risk/sea-risk-data-privacy-in-asean.pdf

having invested in developing frameworks and laws governing the collection, processing and storage of personal data. The economy of monetizing private data, without explicit permission, is being seriously suppressed. The United Nations Development Group (UNDG) has sets out general guidance on data privacy, data protection and data ethics on big data for achievement of the 2030 agenda[7].

While discussing the power and the value of data, it is imperative to discuss the key aspect of how to handle data privacy too. Almost all countries have put forward very clear guidelines, policies or fundamental principles to protect personal data. DLA Piper, a global law firm, has produced an excellent compilation of the various data protection laws around the world[8] (see Figure 3). Their handbook for data protection laws of the world,[9] documents the data protection laws, guidelines or framework in more than a 100 countries worldwide.

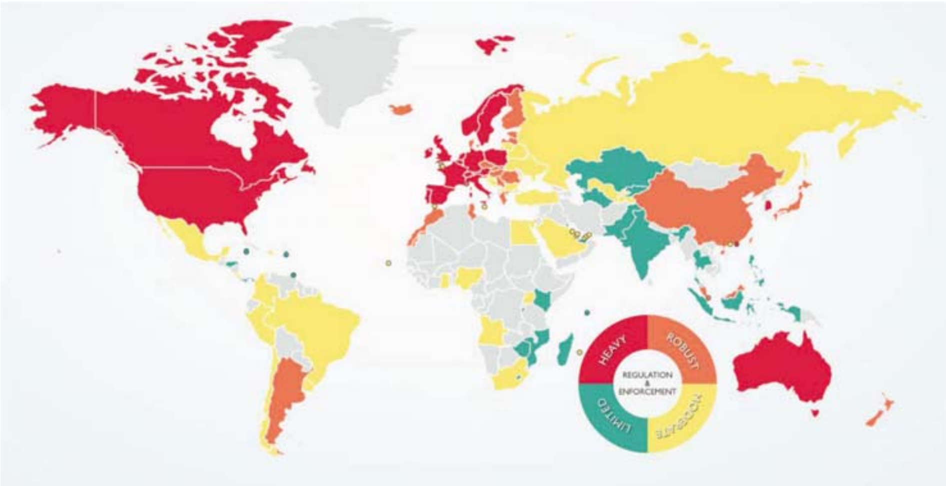


Figure 3: Data protection laws around the world

The General Data Protection Regulation[10] of the European Union outlines very strict regulations on what organizations can do with individual data sets as well as extends the rights of individuals to control how data about them are used. The GDPR proposes the following six data processing/protection principles:

1. *Lawfulness, fairness and transparency:* This emphasizes that the organization's data processing laws are in line with GDPR's requirements and that organizations have a valid reason to collect the data and the data subjects are well aware of the purpose of collecting and processing their data.
2. *Purpose limitation:* This extends the first principle to limiting the use (collecting, processing and storage) of the personal data for the specific purpose for which it was collected.

[7] https://undg.org/wp-content/uploads/2017/11/UNDG_BigData_final_web.pdf
[8] https://www.dlapiperdataprotection.com/
[9] https://www.dlapiperdataprotection.com/#handbook/
[10] https://gdpr-info.eu/

3. *Data minimization*: This clearly states that organizations should collect only the necessary information (adequate, relevant and limited) that is needed to achieve its processing objectives.
4. *Accuracy*: This leads from the principle of data minimization and emphasizes that the personal data collected are kept up to date and adequate steps are taken to ensure the correctness of the data. Data that is not accurate or relevant should be deleted.
5. *Storage limitation:* This stipulates that the personal data must be kept for no longer than is necessary. Data that is no longer required should be deleted because if not, it would conflict with the accuracy principle.
6. *Integrity and confidentiality:* This stipulates protection for personal data in terms of security and confidentiality (against unlawful processing, loss, damage, etc.) using appropriate technical and organizational measures.

These principles coupled with the rights that individuals have to access and control their data makes GDPR one of the most powerful privacy laws today.

Infrastructure and digital security

Infrastructure and digital security is also a major concern. Countries such as India, with the Aadhaar,[11] Gambia with GAMBIS,[12] and the European Union with the proposal to create the Common Identity Repository (CIR) are embarking on creating and maintaining a mega biometrics database of citizens (Cimpanu, 2019). Such a repository is tantalizing for hackers and the huge treasure trove of data would always be a tempting target. Major government databases being hacked are regular news. In 2015, hackers stole some 5.6 million fingerprints from data networks belonging to the government of the United States of America (BBC, 2015). In 2018, a security vulnerability in India's Aadhaar left many people at risk of identity theft (Doshi, 2018). Big data infrastructures are also prone to such vulnerabilities and data breaches; hence, information and infrastructure security should be one of the key components of big data infrastructure management.

Good data versus big data

More data is not always better. In many instances, collecting a selected set of data streams regularly provides more value than aimlessly gathering a broad set of data. Hence, when we talk of big data, we also need to emphasize the need for good data. If analytics were to be built on data that are not reliable or continuous then we would fall into the garbage-in-garbage-out scenario. Analytics plays a key role in realizing the value of the collected data. If effective and efficient algorithms are not designed to separate out actionable intelligence then we are likely to find ourselves drowning in data but thirsting for knowledge.

[11] https://uidai.gov.in/
[12] http://gambis.gm/

Collecting, storing and organizing data

Collecting quality data forms the basic building blocks of data driven agriculture. Initiatives such as the global strategy to improve agricultural and rural statistics[13] help address the declining quantity and quality of agricultural statistics. Open source kits such as KoboToolbox[14] provide simple, robust and powerful tools for data collection. The Computer-Assisted Personal Interview (CAPI)[15] technology assists governments, statistical offices and non-governmental organizations in conducting complex surveys with dynamic structures using tablet devices. The development of CAPI is co-financed by the World Bank, the Bill and Melinda Gates Foundation and the Food and Agriculture Organization of the United Nations. The Open Data Kit[16] community produces free and open-source software for collecting, managing, and using data in resource-constrained environments.

As for storage, in most instances cloud-based storage of big data would be the natural choice as this would facilitate cloud computing and a huge reduction in the cost of operations. Algorithms such as MapReduce are used for processing and generating big data sets with a parallel, distributed computing facility thereby allowing massive scalability. A Hadoop cluster is one example of a computational cluster working in a distributed computing environment that is used to store, analyze and work on huge amounts of unstructured data.

Tools for data analysis, modeling and visualization

To run an optimal analysis and extract insights, it is important that the data is cleaned, sorted and stored in the right way. The majority of a data professional's time should be spent in preparing the data to be in the best shape to facilitate running analytics.

Many tools and platforms are available to prepare the data, transforming and modeling it to facilitate extracting insights. Solutions such as Knime[17] provide an open solution for data-driven innovation. OpenRefine[18] is a helpful tool to work with unstructured data and it can help users explore large data sets with ease. The R-Project[19] supports data mining as well as provides statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. Orange[20] allows open-source machine learning and data visualization.

Data visualization helps in conveying insights generated from big data effectively. For visualization, Power Bi[21] from Microsoft provides interactive visualization and business analytics. Tableau[22] provides an analytics platform to extract insights by facilitating the flow of analysis from data preparation, analytics and collaboration. Infogram[23] helps in creating engaging infographics and reports to help visualize data.

[13] http://www.fao.org/economic/ess/ess-capacity/ess-strategy/en/
[14] https://www.kobotoolbox.org/
[15] http://surveys.worldbank.org/capi
[16] https://opendatakit.org/
[17] https://www.knime.com/
[18] http://openrefine.org/
[19] https://www.r-project.org/
[20] https://orange.biolab.si/
[21] https://powerbi.microsoft.com/en-us/
[22] https://www.tableau.com/
[23] https://infogram.com/

Way forward

Agriculture is increasingly knowledge-intensive. Knowledge derived from combining data from various sources can be used to derive valuable actionable insights. At the farm-level, the farmers of today have to deal with a myriad of data to be able to make livelihood-based decisions on a regular basis. Data on soil health, weather, irrigation, markets, early-warning systems, diseases and pests, finance/loan availability, as well as government-related information/subsidies all come into play in making a decision at the farm-level. At the province or district level, the policymaker has to have real-time or near real-time information on market prices, projected yield of a particular crop at the end of the harvest season, beneficiaries of government schemes/subsidies, efficacy of pre-emptive actions to protect against diseases and pests, disaster mitigation and much more. At the national level, quality data would help design effective policies to assist smallholder farmers, monitor and remove inefficiencies in the value-chains, ensure consumers about quality produce and to eradicate hunger, malnutrition and ensure food security in the country. At the global level, a coalition including FAO, the Bill and Melinda Gates Foundation and national governments has launched a USD 500 million effort to help developing countries gather data on small-scale farmers to help fight hunger and promote rural development (Tollefson, 2018). Agriculture, as other fields, will have the unprecedented capability to extract intelligence and make evidence-based decisions firmly grounded on real-time, reliable data and effective analytics. Introduction of new actors in the value chain and investment in developing the capacities of existing actors will be some of the key challenges that governments will have to address. Newer ways of harvesting data will lead to granular insights that previously was impossible to achieve. The UN Global Pulse recently explored using data from social media and public broadcasters to extract insights to feed early warning systems and to inform peace and security processes in Africa (Hidalgo-Sanchis, 2018).

We should not miss the opportunity to harness actionable intelligence from big data to achieve the Sustainable Development Goals. The cost of inaction would be far greater if we do not put in place an organizational data ecosystem to utilize the wealth of data that we have and to derive valuable insights from it.

References

Ali, A., Qadir, J., Rasool, R., Sathiaseelan, A., Zwitter, A. & Crowcroft, J. 2016. Big data for development: applications and techniques, *Big Data Analytics* 1: 2. (Also available at https://doi.org/10.1186/s41044-016-0002-4).

BBC. 24 September 2015. Millions of fingerprints stolen in US government hack [online]. [Cited 15 February 2019]. https://www.bbc.co.uk/news/technology-34346802

Cimpanu, C. 22 April 2019. EU votes to create giant biometrics database [online]. [Cited 28 April 2019]. https://www.zdnet.com/article/eu-votes-to-create-gigantic-biometrics-database/

DLAPiper. 2014. Data protection laws of the world handbook: third edition [online]. [Cited 14 February 2019]. www.dlapiperdataprotection.com/#handbook

Doshi, V. 2018. A security breach in India has left a billion people at the risk of identity theft. *Washington Post* 4 January 2018 (also available at https://www.washingtonpost.com/news/worldviews/wp/2018/01/04/a-security-breach-in-india-has-left-a-billion-people-at-risk-of-identity-theft/?utm_term=.fc94931e4afc).

Economist. 2017. The world's most valuable resource is not oil, but data. Print edition 6 May 2017 (also available at https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data).

Hidalgo-Sanchis, P. 2018. Using big data to support peace and security efforts in Africa. United Nations Global Pulse. Blog 26 December 2018 [online]. [Cited 14 February 2019]. https://www.unglobalpulse.org/news/using-big-data-and-ai-support-peace-and-security-efforts-africa

Holak, B. 2019. Demand for data scientists is booming and will only increase. TechTarget. Search Business Analytics 31 January 2019 [online]. [Cited 22 March 2019]. https://searchbusinessanalytics.techtarget.com/feature/Demand-for-data-scientists-is-booming-and-will-increase

Marr, B. 2018. How much data do we create every day? The mind-blowing stats everyone should read. *Forbes* 21 May 2018 [online]. [Cited 24 February 2019]. https://www.forbes.com/sites/bernardmarr/2018/05/21/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/#7e1ce72e60ba

Reid, D. 2017. Mastercard's boss just told a Saudi audience that 'data is the new oil' CNBC News 20 October 2017 [online]. [Cited 2 March 2019]. https://www.cnbc.com/2017/10/24/mastercard-boss-just-said-data-is-the-new-oil.html

Tollefson, 2018. Big data project aims to transform farming in world's poorest countries. News. *Nature: International Journal of Science* 24 September 2018 [online]. [Cited 4 February 2019]. https://www.nature.com/articles/d41586-018-06800-8

Author

Gerard Sylvester

Gerard.Sylvester@fao.org

Food and Agriculture Organization of the United Nations

1 NO POVERTY
Spending patterns on mobile phone services can provide proxy indicators of income levels

2 ZERO HUNGER
Crowdsourcing or tracking of food prices listed online can help monitor food security in near real-time

3 GOOD HEALTH AND WELL-BEING
Mapping the movement of mobile phone users can help predict the spread of infectious diseases

4 QUALITY EDUCATION
Citizen reporting can reveal reasons for student drop-out rates

5 GENDER EQUALITY
Analysis of financial transactions can reveal the spending patterns and different impacts of economic shocks on men and women

6 CLEAN WATER AND SANITATION
Sensors connected to water pumps can track access to clean water

7 AFFORDABLE AND CLEAN ENERGY
Smart metering allows utility companies to increase or restrict the flow of electricity, gas or water to reduce waste and ensure adequate supply at peak periods

8 DECENT WORK AND ECONOMIC GROWTH
Patterns in global postal traffic can provide indicators such as economic growth, remittances, trade and GDP

9 INDUSTRY, INNOVATION AND INFRASTRUCTURE
Data from GPS devices can be used for traffic control and to improve public transport

10 REDUCED INEQUALITY
Speech-to-text analytics on local radio content can reveal discrimination concerns and support policy response

11 SUSTAINABLE CITIES AND COMMUNITIES
Satellite remote sensing can track encroachment on public land or spaces such as parks and forests

12 RESPONSIBLE CONSUMPTION AND PRODUCTION
Online search patterns or e-commerce transactions can reveal the pace of transition to energy efficient products

13 CLIMATE ACTION
Combining satellite imagery, crowd-sourced witness accounts and open data can help track deforestation

14 LIFE BELOW WATER
Maritime vessel tracking data can reveal illegal, unregulated and unreported fishing activities

15 LIFE ON LAND
Social media monitoring can support disaster management with real-time information on victim location, effects and strength of forest fires or haze

16 PEACE, JUSTICE AND STRONG INSTITUTIONS
Sentiment analysis of social media can reveal public opinion on effective governance, public service delivery or human rights

17 PARTNERSHIPS FOR THE GOALS
Partnerships to enable the combining of statistics, mobile and internet data can provide a better and real-time understanding of today's hyper-connected world

Source: Data Privacy, Ethics and Protection: Guidance Note on Big Data for Achievement of the 2030 Agenda. https://undg.org/wp-content/uploads/2017/11/UNDG_BigData_final_web.pdf

How data analytics can support the SDGs

Article

2

Big data: a shift in paradigm towards digital agriculture

Today, agriculture faces urgent challenges of feeding a growing global population (forecast to reach 8.5 billion people by 2030) in the face of diminishing arable areas, increasingly complex supply chains, accelerating climate change and increased uncertainties. There is evidence to suggest that for staple cereal crops, crop yields might decline by 10 percent, or even as high as 15 percent to 17 percent for every degree of temperature rise (Wallace-Wells, 2017). A shift to digital agriculture offers opportunities to help increase crop yields, reduce food losses and make agricultural supply chains more efficient as well as improve food distribution and retail, which are often cited as major problems.

In essence, digital agriculture relies on quality data to gather information, improve decision-making, enable innovative services and enhance communication amongst agriculture sector stakeholders. Over the years, the role of information and communications technologies (ICTs) has evolved from using telephones, television, radio, computers and Internet for end-user communication to using sensors and data analytics to help drive precision agriculture, improve yields, of supply chains, solutions and management.

The quality, granularity and variety of information can help make the agricultural sector more efficient, as well as enable innovative services in related sectors such as payments, insurance, credit, workforce management, logistics and public subsidies (Figure 4). The availability of these datasets to third parties under license or via open data, depending on the type of data and national requirements, could potentially accelerate innovation, promote job creation and incentivize digital skills.

Today, a growing number of data sources are helping drive innovation in digital agriculture services and solutions. Sensors in fields are starting to provide a wide range of granular data points on soil conditions, as well as granular information on wind, fertilizer requirements, water availability and pest infestations. GPS units on tractors, can help optimize usage of heavy equipment (Schriber, no date) (Table 1).

Data analytics can help prevent spoilage by moving products faster and more efficiently through supply chains. Unmanned aerial vehicles (UAVs) and drones can patrol and monitor fields and alert farmers to crop ripeness or potential problems. Radio frequency identity (RFID)-based traceability systems follow farm products as they move through the supply chain, from the farm to the compost or recycle bin (Wang *et al.*, 2006). Individual crops can be monitored for nutrients and growth rates. Data analytics can help farmers determine the best crops to plant, in view of profitability and sustainability. Agricultural technology can also help farmers hedge against losses and even out cash flows (Sparapani, 2017).

Source: FAO, ITU.

Figure 4: Changing role of digital technologies in agriculture

Table 1: Agricultural uses of tools and sensors

Purpose	Example
Camera	Provides pictures of leaf health, lighting brightness, chlorophyll measurement and ripeness level. Also used for measuring leaf area index (LAI) and measuring soil organic and carbon lmake-up.
Global positioning system (GPS)	Provides location for crop mapping, disease/pest location alerts, solar radiation predictions and fertilizing.
Microphone	Helps with predictive maintenance of machinery.
Accelerometer	Helps determine leaf angle index, and used as an equipment rollover alarm.
Sensors	Temperature, UV exposure and humidity in crops.Temperature, pressure, sound in equipment to predict upcoming failure.
Gyroscope	Detects equipment rollover.

Source: Schriber (no date)

In addition, ICTs and data can enhance decision-making throughout the broader supply chain with features such as:

- geolocation and land use (e.g. GIS maps, satellite images);
- digital identity and individual linked data (e.g. agriculture subsidies);
- financial services (e.g. banking details, digital payments);
- communications (e.g. telephone, email, television, radio, instant messaging, web pages, social media);
- agricultural statistics and databases;
- agricultural market related data (e.g. prices, supply, demand, export/import regulations);
- advisory, awareness and capacity-building (e.g. good agricultural practices content);
- logistics data (e.g. transport, cold storage availability and location);
- data relating to the management of disasters and/or pests (e.g. pests and disease alert and analysis, weather-related disasters);
- provenance and trust across the value chain (e.g. traceability using blockchain); and
- passive observation of herds and livestock, to monitor their health and wellbeing, identify sick individuals and potentially even trigger treatments using artificial intelligence (AI) and machine learning (ML) models.

The size, speed and complexity of datasets have rendered the traditional analytical tools outdated and created a new paradigm of big data (Text Box 1). "Open data" is another movement, referring to publicly accessible data that people, companies, and organizations can use to launch new ventures, analyze patterns and trends, make data-driven decisions, and solve complex problems. Technically, this requires:

Text Box 1: Big data: definition and characteristics

Definition: "A paradigm for enabling the collection, storage, management, analysis and visualization, potentially under real-time constraints, of extensive datasets with heterogeneous characteristics."

Characteristics:

Volume	Variety	Velocity
amount of data collected, stored, analysed and visualized, which big data technologies need to resolve	different data types and data formats that are processed by big data technologies	refers to both how fast the data is being collected and how fast the data is processed by big data technologies to deliver expected results

Additionally, some commentators also add "veracity", which refers to the uncertainty of the data, and "value", which refers to the business results from the gains in new information using big data technologies. Other characteristics may also be added.

Source: ITU-T Recommendation Y.3600

- data publication – metadata supporting machine readability, data format, and licenses;
- data finding – data identification, data semantics, and data access; and
- data provenance[24] – data quality, data lineage tracking, checking, verification and data versioning.

Big data is a paradigm change that has impact across the SDGs (see page 10). The ability to collect, collate and use big data to build advanced models for crop yields or even machine learning (ML) of crop yields, particularly in light of global warming, is another step. Advanced computer simulations have been used to model and map the impact of climate change, to map poverty in rural districts of Mexico, to map electrification of rural areas in India, the yields of crops, as well as deforestation in the Amazon Basin and Malaysia.

Sharing and interoperability of data and information systems are key necessities to harness the data collected (Figure 5). Countries have adopted e-government interoperability frameworks to enable better sharing and interoperability. Afghanistan, Australia, Bhutan, Hong Kong (China), India, New Zealand are some examples from Asia-Pacific economies that have adopted e-government interoperability frameworks. Interoperability and reuse of valuable, high-variety, and high-volume data from the agricultural and non-agricultural sectors remain a key challenge, which has also been experienced in FAO and ITU's experiences in supporting countries in developing e-agriculture strategies.

In the fight against climate change, big data and open data can also be used to help green urban areas and encourage parks and plantations within the limits of expanding urban centres. For example, trees.sg has compiled a database of 500 000 urban trees, cataloguing species, health, flowering and even pruning schedules. This will enable the City of Singapore to monitor tree growth/loss and to involve citizens in tree upkeep (Gov Tech Singapore, 2018).

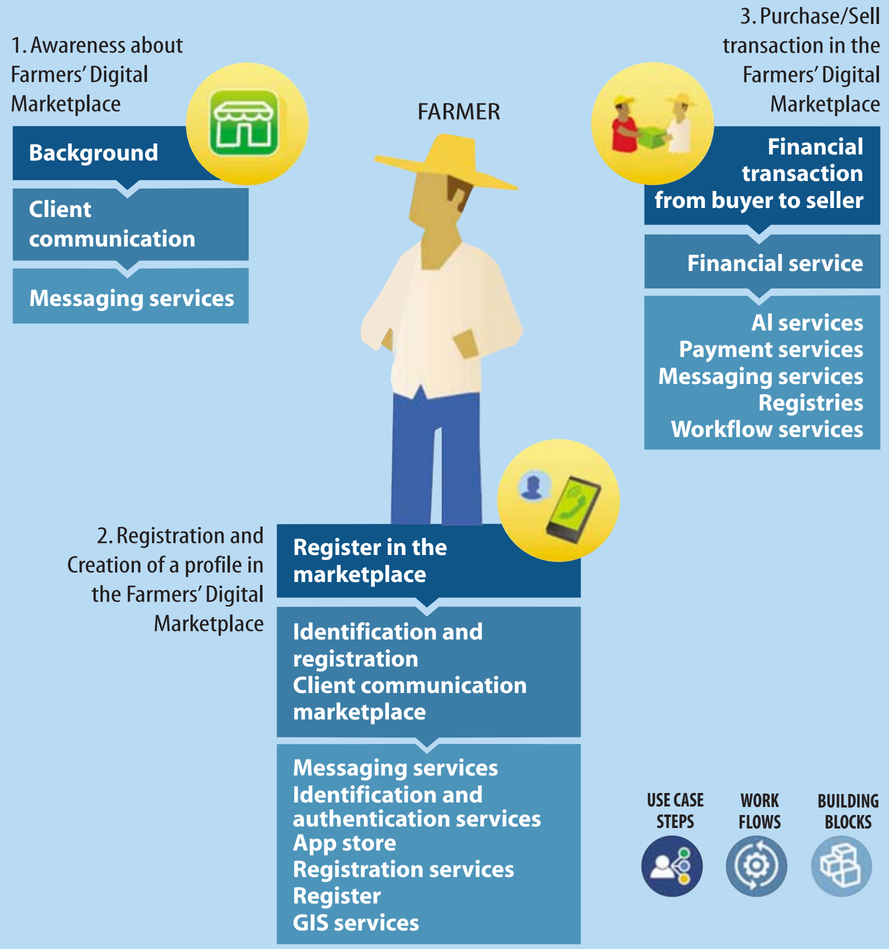
[24] Big data provenance: information that records the historical path of data according to the data lifecycle operations in a big data ecosystem. Data lifecycle operations include data generation, transmission, storage, use, and deletion. Data provenance information provides details about the source of data, such as the person responsible for the provision of data, functions applied to data, and information about the computing environment. For details, please see ITU, 2018c.

Summary: Connecting rural farmers to market information, products, and related services to improve rural incomes.

Sector: Agriculture

Mapped SDG Targets:

- 2.1 By 2030, end hunger and ensure access by all people, in particular the poor and people in vulnerable situations, including infants, to safe, nutritious and sufficient food all year round
- 2.2 By 2030, end all forms of malnutrition, including achieving, by 2025, the internationally agreed targets on stunting and wasting in children under 5 years of age, and address the nutritional needs of adolescent girls, pregnant and lactating women and older persons
- 2.3 By 2030, double the agricultural productivity and incomes of small-scale food producers, in particular women, indigenous peoples, family farmers, pastoralists and fishers, including through secure and equal access to land, other productive resources and inputs, knowledge, financial services, markets and opportunities for value addition and non-farm employment



Source: ITU, 2019

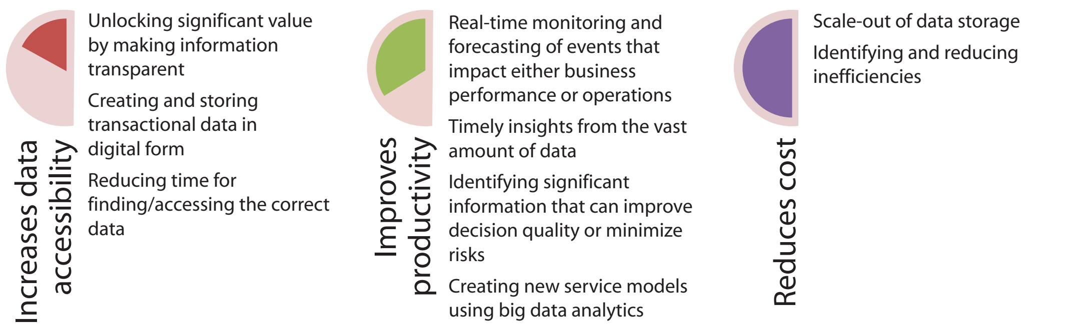
Figure 5: Farmers' digital marketplace for market linkage
Raghu, a smallholder farmer, uses various digital systems as part of a broader digital market place

Which benefits do big data technologies offer for digital agriculture?

Big data technologies and services offer significant benefits (Figure 6) and can address challenges of heterogeneity and incompleteness of data, need for processing of large and rapidly increasing volumes, timeliness requirements, and privacy concerns amongst others.

Big data also drives use of new technologies such as AI, ML, cloud computing, distributed ledger technologies (DLT) (e.g. blockchain), and the Internet of things (IoT). For example:

- Various IoT-based applications will improve agricultural processes and make farms smarter, such as precision farming, driverless tractors, agricultural drones/robots, monitoring, smart irrigation and smart greenhouses (Huawei, no date).
- Smart and precision agriculture applications use artificial intelligence (AI) to interpret data captured by other machines (e.g. drones or sensors) and create models and strategic inferences (e.g. whether or not to initiate harvesting or pesticide spraying) and activate other smart machines (e.g. a remote operated pump-set to start irrigation of fields, farm robots or self-driven tractors).
- Machine vision (image recognition) is used for diagnosing pests or soil defects (ITU, 2019).



Challenges are important considerations for big data exchange	Big data ecosystems areexpected to provide the following benefits:
Various sources, types and formats of data: big data service providers have to handle diverse aspects of data and data sources during data collection, storage and integration. **Schema-on-read:** usually big data is stored in a raw format, but after data is discovered and captured, it is transformed to fulfil the application's requirements. **Unawareness of suitable data/Unconstrained usage of data:** sometimes, the big data service customer does not recognize what kinds of data are really needed. This often leads to unconstrained usage of data in a big data ecosystem.	**mitigation of silos in the ecosystem** through a better sharing of high-variety data between involved parties; **monetization of data** enabling better revenues to be made by parties from high volume of data exchanged in the ecosystem; **openness of the publicly available data** contributing to human society and economic activities; **facilitation of the appearance of new and effective business models**; and **interconnection of valuable, high-variety, and high-volume data** contributing more to human society and economic activities.
Source: ITU, 2018b	

Source: ITU. 2015

Figure 6: Big data benefits and challenges

The amount of data generated by the average farm each day is increasing dramatically (Meola, 2016) and shows the potential of AI in this domain. For instance, it has been suggested that algorithms could identify 26 plant diseases in 14 different species with 99 percent accuracy (Mohanty *et al*., 2016).

The integration of big data with cloud computing services provides further advantages of **scalability** (allows the big data service user to easily upscale or downscale the resources quickly); **resiliency** (helps maintain an acceptable level of service in the face of faults affecting normal operations of systems); **cost effectiveness** (lowering costs of storage and analysis); **efficient analysis**; and **deep information extraction.**

Who are the key actors in the big data agricultural ecosystem?

The big data ecosystem (Figure 7) comprises data providers (including data suppliers and data brokers), big data service providers and big data service customers.

Data suppliers provide data from different sources to the data broker and this can be accessed by the big data service providers.

Data brokers connect between the data supplier and the big data service provider. The data broker can act as a clearinghouse, open data market, etc.

Big data service providers support capabilities for big data analytics and infrastructure. Big data service providers can act as a big data platform, an extension of the existing data analytics platform, etc.

Big data service customers are the end-users or a system that uses the results or services from a big data service provider. Big data service customers may produce new services or knowledge on consumer activities and furnish them outside the big data ecosystem.

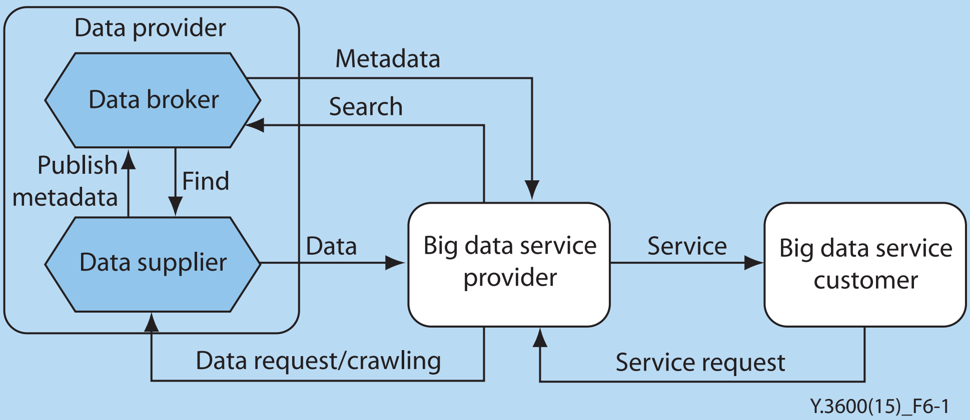
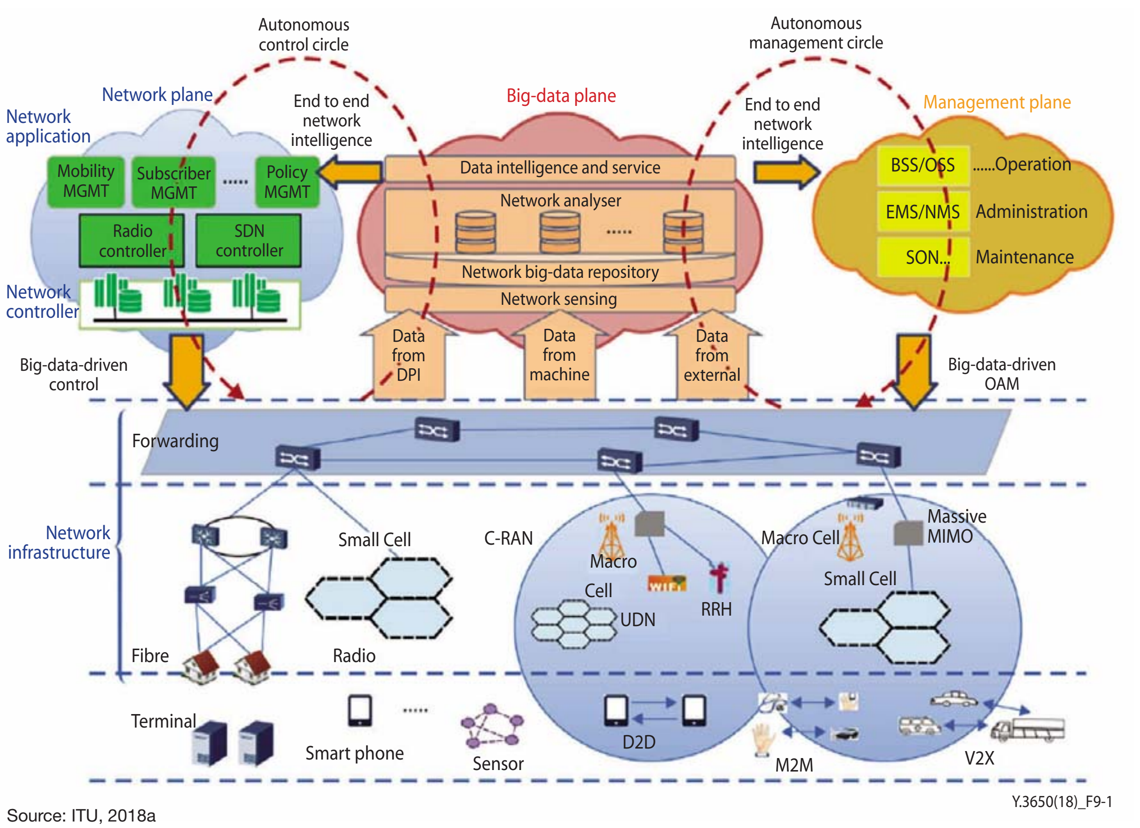


Figure 7: Big data stakeholders

Delivering big data over networks

The real value of big data is realized when information is collected from various sources (e.g. computer terminals and servers, smartphones, sensors, devices, machine, vehicles) and transported using a telecommunication network infrastructure (SMS, fiber, radio, mobile, copper, satellite), stored in the cloud and potentially shared across various services. The information gathered can also be used to improve the performance of the network itself and provide valuable insights to its users (Figure 8).



Source: ITU, 2018a

Figure 8: Big data over networks

From an end user's perspective, mobile phones and tablets, and the services rendered over them are of prime interest. At the end of 2018, about 96 percent of the global population lived within reach of a mobile-cellular network[25] and there were some 5.3 billion unique mobile subscribers globally. These networks and devices are designed to provide faster broadband access and have created what is now called the "app economy". The next generation of mobile networks, termed 5G (or IMT-2020) networks, are expected to cater to enhanced mobile broadband (more than 10 Gbps), as well as demonstrating features of massive machine type communication (greater than 1 million devices per square kilometre) and ultra-reliable and low latency communication (at 1 millisecond or less).

[25] ITU's Connect 2030 agenda targets 96 percent of the world population covered by broadband services by 2023.

In the agricultural sector, 5G and advanced connectivity could enable a range of new innovative services utilizing features such as sensing, logistics, smart agriculture, industrial automation, remote diseases detection, augmented and virtual reality (VR). Big data capabilities and a well-developed ecosystem underpin digital agriculture. Standardization is a key driver to enable this collaborative environment and develop the economies of scale and scope envisioned.

Big data standardization efforts

Standardization requires stakeholder coordination (public and private) at both the international and national levels. However, it is important to note that standardization is an ongoing activity and is operationalized only when adopted either mandatorily (by national standardization organizations or organizational policies and regulations) or voluntarily.

In the big data landscape, standardization development organizations (SDOs) including ITU-T (SG 13, 17 and 20), International Organization for Standardization (ISO/IEC JTC 1), World Wide Web Consortium (W3C), Organization for the Advancement of Structured Information Standards (OASIS) technical committees (TCs), Data Mining Group (DMG), TM Forum (formerly TeleManagement Forum) have made significant progress in this regard.

An insight into the dimensions of standardization areas and gap analysis can be thought of as a matrix composed of two axes (Table 2) including:

- subject of applications including general definition: the standard which provides general descriptions or terms and definitions of the technology; common requirements, use cases: architecture; API, interface, profile; data model, format, schema; others (e.g. guidelines, technical reports); and
- related technologies for supporting big data including fundamental concept of big data and its applications; data exchange; data integration; analysis/visualization; data provenance/metadata; security/privacy for big data, especially personal identification information; other big data related technologies.

Furthermore, the agricultural sector is characterized by heterogeneous machinery and diverse process partners. Problems arise as a result of idle times in agricultural processes, sub-optimal machine utilization and incorrect planning. Other problems are caused by incompatibilities of machines from different manufacturers. A standardized communication language is required to support communication between heterogeneous machines.

Table 2: Standardization matrix of big data[26]

	General/ Definition	Common requirement/Use case	Architecture	API, Interface and its profile	Data model, format, schema	Others (e.g., guideline)
Fundamental	**ITU-T Y.3600** ISO/IEC 20546 ISO/IEC 20547-1	**ITU-T Y.3600**	**ITU-T Y.3519** **ITU-T Y.BD-arch** ISO/IEC 20547-3			
Data exchange, Data flow	**ITU-T Y.3601** ISO/IEC 19944 ISO/IEC 22624 ISO/IEC 23751	**ITU-T Y.3601** ISO/IEC 22624 ISO/IEC 23751			OASIS AMQP 1.0 OASIS MQTT 3.1.1	
Data integration	**ITU-T Y.bdi-reqts**	**ITU-T Y.bdi-reqts**			W3C DCAT W3C JSON-LD 1.0 W3C LDP 1.0 W3C RDF 1.1 W3C OO	
Analysis/ Visualization				DMG PFA	DMG PMML 4.3 DMG PFA	TMF BDAG
Data Provenance/ Metadata/Trust	**ITU-T Y.3602**	**ITU-T Y.3602** ISO/IEC TR 23186			**ITU-T Y.bdm-sch** W3C MVTD W3C MTDMW	
Security/Privacy Data protection	**ITU-T X.1601** **ITU-T D.princip_bigdata** **ITU-T X.mdcv** ISO/IEC 27000 IEO/IEC 29100	**ITU-T X.1147** ISO/IEC 20547-4 ISO/IEC 27001			ISO/IEC 27002 ISO/IEC 27018	**ITU-T X.1641** **ITU-T X.GSBDaaS** **ITU-T X.sgBDIP** **ITU-T X.sgtBD**
Application		**ITU-T F.VSBD** ISO/IEC 20547-2	**ITU-T H.VSBD**			
Network and infrastructure	**ITU-T Y.bDPI-Mec** **ITU-T Y.3650**	**ITU-T Y.IoT-4114** **ITU-T Y.Sup.50** **ITU-T Y.bDDN-req** **ITU-T Y.BDDP-reqts** **ITU-T F.AFBDI** **ITU-T Y.3505**	**ITU-T Y.3302** **ITU-T Y.bDDN-FunArch**			**ITU-T Y.3651**
Others						**ITU-T Study_ bigdata** ISO/IEC 20547-5

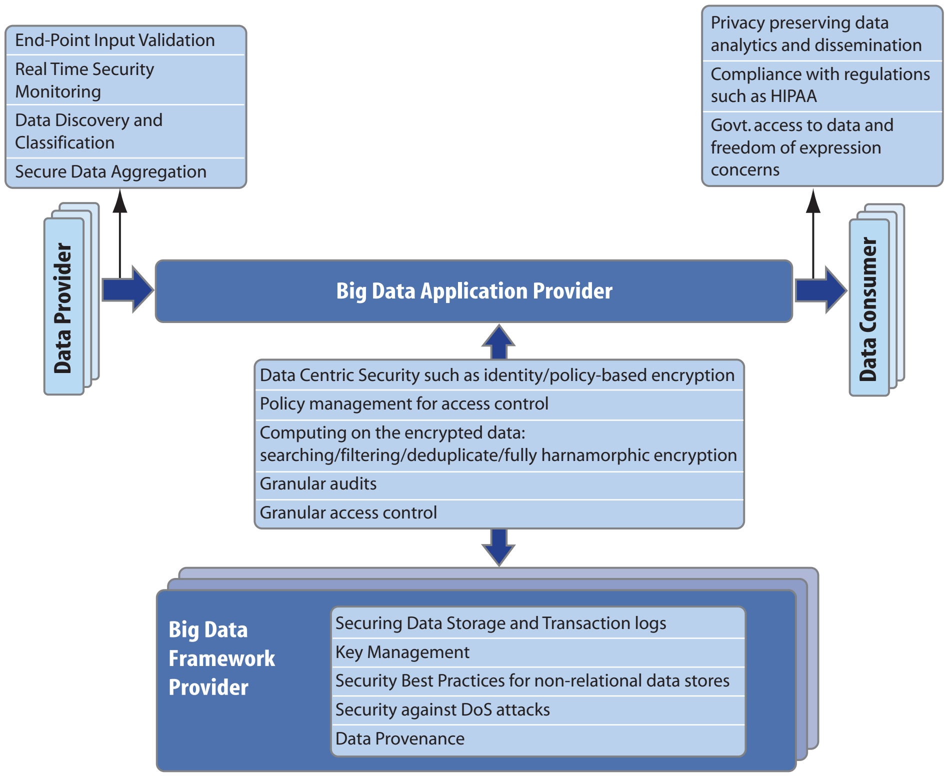
[26] Draft revised Supplement ITU-T Y.sup.bdsr2: Supplement on Big data Standardization roadmap (https://www.itu.int/md/T17-SG13-190304-TD-WP2-0394/en)

Security and privacy in big data

Security and privacy requirements are amplified in the big data environment, largely resulting from its distributed nature and advanced technical characteristics. Data come from different industries, different types of sources, are varying in size and are subject to different security and legal environments. For example, data such as national biometric IDs and individual profiles are subject to data privacy laws in some countries, and are getting greater regulatory attention at the international and national level. In the new environment, there is an increasing reliance on IoT-based sensing, which creates new vulnerabilities.

Big data also attracts greater attention owing to its high value. Some identified challenges concerning big data in the context of security and privacy are: secure computations in distributed programming frameworks; secure data storage and transactions logs; end-point input validation/filtering and data provenance; real-time security/compliance monitoring; scalable and composable privacy-preserving data mining and analytics; and anonymization and de-identification.

The National Institute of Standards and Technology (NIST), United States of America, provides an overview of several security and privacy topics (Figure 9) with respect to some key NIST Big Data Reference Architecture (NBDRA) components and interfaces.



Source: Adapted from NIST, 2015

Figure 9: National security and privacy fabric overlay to the NBDRA

New data driven e-agriculture solutions expose the stakeholders to new risks. A report of the United States Department of Homeland Security (DHS, 2018) identifies the key threats to precision agriculture using the “confidentiality, integrity and availability” model of information security. For example, intentional falsification of data can disrupt crop or livestock sectors, or the introduction of rogue data into a sensor network can damage a crop or herd.

It is important that digital agricultural solutions respect and adapt to the national privacy and security frameworks in force. For example, the European Union introduced the General Data Protection Regulation (GDPR) in force since 25 May 2018, and farming applications in Europe or relating to European citizens may have to comply with the GDPR. A framework on personal data protection was adopted by ASEAN Telecommunications and Information Technology Ministers (TELMIN) in 2016 (ASEAN Secretariat, 2016) and in 2018 the ASEAN TELMIN endorsed a framework on digital data governance (ASEAN Secretariat, 2018).

With regards to digital agriculture, some countries may not wish certain data or information on their national food supply or agricultural risks and vulnerabilities to become common knowledge through open data portals or may require those data to be stored locally. The data may also be exploited in different ways by different stakeholders. For example, satellite images and ML models of droughts or crop yields in sub-Saharan Africa are used by some commercial commodity houses to enable them to take positions for profit on the commercial futures and options markets.

Conclusions

Big data – and the ability to use big data to model farming – are creating and spreading a new technology paradigm for innovation in agriculture. The ability to use interoperable massive and diverse data sources offers considerable promise for agriculture and the lives of its stakeholders, especially given the additional uncertainties created by climate change and global warming. Big data also brings new actors into the ecosystem and can create positive secondary effects in other sectors such as banking, insurance, logistics, and public services. However, governments need to work together with agricultural stakeholders to create an enabling policy environment that fosters the ecosystem and addresses the privacy and security concerns.

References

ASEAN Secretariat. 2018. *Framework on digital data governance*. ASEAN Telecommunications and Information Technology Ministers (TELMIN) meeting [online]. Jakarta. [Cited 24 May 2019]. https://asean.org/storage/2012/05/6B-ASEAN-Framework-on-Digital-Data-Governance_Endorsedv1.pdf

ASEAN Secretariat. 2016. *Framework on personal data protection*. ASEAN Telecommunications and Information Technology Ministers (TELMIN) meeting [online]. Jakarta. [Cited 24 May 2019]. https://asean.org/wp-content/uploads/2012/05/10-ASEAN-Framework-on-PDP.pdf

Department of Homeland Security. United States of America. 2018. *Threats to precision agriculture* [online]. United States of America. [Cited 24 May 2019]. https://www.dhs.gov/sites/default/files/publications/2018%20AEP_Threats_to_Precision_Agriculture.pdf

Gov Tech Singapore. 2018. *The inside story of how NParks mapped 500,000 trees in Singapore on trees.sg* [online]. Singapore. [Cited 24 April 2019]. https://www.tech.gov.sg/media/technews/the-inside-story-of-how-nparks-mapped-500000-trees-in-singapore-on-treessg?utm_source=newsletter&utm_medium=email3&utm_campaign=23Apr19

Huawei. no date. *The connected farm: a smart agriculture market assessment* [online]. India. [Cited 24 April 2019]. https://www.huawei.com/-/media/CORPORATE/Images/PDF/v2-smart-agriculture-0517.pdf?la=en

ITU. 2019. *SDG digital investment framework: a whole government approach to investing in digital technologies to achieve the SDGs*. Geneva, International Telecommunication Union. (Also available at https://www.itu.int/dms_pub/itu-d/opb/str/D-STR-DIGITAL.02-2019-PDF-E.pdf).

ITU. 2018a. Series Y: Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of things and smart cities. Cloud computing. *Framework of big-data-driven networking. Recommendation ITU-T Y.3650* [online]. [Cited 23 May 2019]. https://www.itu.int/rec/T-REC-Y.3650-201801-I/en

ITU. 2018b. Series Y: Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of things and smart cities. Cloud computing. *Big data – Framework and requirements for data exchange. Recommendation ITU-T Y.3601* [online]. [Cited 23 May 2019]. https://www.itu.int/rec/T-REC-Y.3601-201805-I/en

ITU. 2018c. Series Y: Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of things and smart cities. Cloud computing. *Big data – Functional requirements for data provenance. Recommendation ITU-T Y.3602* [online]. [Cited 23 May 2019]. https://www.itu.int/rec/T-REC-Y.3602-201812-I

ITU. 2015. Series Y: Global information infrastructure, Internet protocol aspects, next-generation networks, Internet of things and smart cities. Cloud computing. *Big data – Cloud computing based requirements and capabilities. Recommendation ITU-T Y.3600* [online]. [Cited 23 May 2019]. https://www.itu.int/rec/T-REC-Y.3600-201511-I

Meola, A. 2016. Why IOT, big data and smart farming are the future of agriculture. *Business Insider* 26 December 2016 [online]. [Cited 25 April 2019]. https://www.businessinsider.com/internet-of-things-smart-agriculture-2016-10

Mohanty, S.P., Hughes, D. & Salathe, M. 2016. *Using deep learning for image-based plant disease detection* [online]. United States of America. [Cited 24 May 2019]. https://arxiv.org/pdf/1604.03169.pdf

The National Institute of Standards and Technology (NIST). 2015. *NIST big data interoperability framework: volume 4, security and privacy*. NIST Special Publication 1500-4 [online]. United States of America. [Cited 26 May 2019]. https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1500-4.pdf

Schriber, S. no date. *Smart agriculture sensors: helping farmers and positively impacting global issues, too* [online] United States of America. [Cited 23 April 2019]. https://www.mouser.ch/applications/smart-agriculture-sensors/

Sparapani, T. 2017. How big data and tech will improve agriculture, from farm to table. *Forbes*, 23 March 2017 [online]. United States of America. [Cited 23 April 2019]. https://www.forbes.com/sites/timsparapani/2017/03/23/how-big-data-and-tech-will-improve-agriculture-from-farm-to-table/#54bffa505989

Wallace-Wells, D. 2017. The uninhabitable earth. *New York Magazine* 10 July 2017. (Also available at http://nymag.com/intelligencer/2017/07/climate-change-earth-too-hot-for-humans.html).

Wang, N., Zhang, N. & Wang, M. 2006. Wireless sensors in agriculture and food industry–recent development and future perspective. *Computers and Electronics in Agriculture*, 50(1): 1–14. DOI: 10.1016.j.compag.2005.09.003.

Author

Ashish Narayan (ashish.narayan@itu.int), **Hani Eskandar** (hani.eskandar@itu.int) **and Phillippa Biggs** (phillippa.biggs@itu.int)
International Telecommunication Union (ITU)

Article

3

Unpacking the data driven digital revolution

The digital revolution, aided by technological advancements, especially in processes that capture/deliver, store, access, and analyze data, is driving the development of information societies, that is, societies in which essential services are underpinned largely by data. It has become the new normal for many people, for example, to check on traffic updates and the most optimal route on Google Maps, before stepping out of their homes. Intelligent solutions such as Google Maps are accessible to literally anyone with an Internet connection. Moreover, such intelligent solutions are also starting to make an impact on the lives of smallholder farmers, who face a number of serious risks, such as those posed by climate, the need to obtain financing. Smallholder farmers are also starting to reap the benefits of technology. The example of rice-producing farmers in Colombia is instructive. Rice yields across the country have been on the decline from 6 tonnes to 5 tonnes per hectare between 2005 and 2013. Experts ascribe most of this variation to fluctuating weather patterns, which can reduce yields by between 30 percent and 40 percent. Part of the problem (and the opportunity) in Colombia was that farmers did not have timely access to weather forecasts and intelligent decision support systems. A group of scientists from the Climate Change Agriculture and Food Security (CCAFS) programme at the International Center for Tropical Agriculture, teamed up with the national agricultural extension system and the rice cooperative FEDEARROZ and analyzed the complex relationships between the historical climate and the crop management practices of these farmers. After obtaining new insights, they used the science of weather forecasts to produce a decision support system that was accessible to the farmers through the rice cooperative. The insight at the start of 2014 was that the first planting season of the year was not conducive to optimal yields, but the second one was. The cooperative issued an advisory note suggesting that farmers should not plant rice in the first season of 2014. Approximately 170 farmers followed the advice and the cooperative was able to avoid losses of close to USD 1.7 million (CCAFS, 2014). This is one of the many examples of how data and digital technologies are beginning to revolutionize key and essential services for poor and small farmers, who contribute from 30 percent to 34 percent of the food produced worldwide (Ricciardi *et al.*, 2018).

This document assesses how data and digital technologies are transforming farming-related livelihoods through smart and intelligent digital solutions for stakeholders all across the agricultural value chain, and are reaching scale. The smallholder farmer data ecosystem, which consists of processes and stakeholders that capture, store, analyze and develop new services, is witnessing the emergence of digital profiles, data warehouses and data lakes. These enable users to improve their understanding of the value chain via high resolution spatio-temporal monitoring of almost every process in the value chain, from production to sale. In addition, the benefits of merging multiple data sources, stored in data lakes and data warehouses, with advanced analytics to generate intelligent and smart services for stakeholders along the value chain are also becoming evident. This document also attempts, to characterize the factors that are contributing to digital transformation across the smallholder farmer-associated agricultural value chain.

One challenge is that there is a variety of solution providers in the data ecosystem, each with their own, non-interoperable data warehouse and data lakes. We foresee an opportunity for the development of much smarter, advanced monitoring and decision support systems based on making these diverse data warehouses/data lake systems findable, accessible, interoperable, and reusable (FAIR).

The status and constraints faced by smallholder farmers and their value chain

Smallholder farmers are defined by the Food and Agriculture Organization of the United Nations (FAO) as those who own less than two hectares of land (Lowder *et al.*, 2016). It estimates that about 570 million smallholders manage anywhere between 12 percent and 24 percent of the global agricultural area; however, a more recent study has shown that roughly 40 percent of the global agricultural area is managed by smallholders (Lesiv *et al.*, 2019). Regardless of these differing estimates, a large proportion of the world's population is employed in smallholder farming and smallholders produce an equally large proportion of the food consumed globally. The irony is that, despite their importance to the world population's well-being, smallholders remain at the bottom of the economic pyramid, living on less than two dollars a day (The World Bank, 2016) and their families are often severely malnourished (FAO, 2014). This disparity also has a bearing on the food system – sustainable food production, food availability and access – as there is significant evidence of increased internal migration (rural-rural or rural-urban relocation) by smallholder farming communities. Such internal migration is driven by farmers' inability to adapt to stresses and shocks (e.g. biophysical and financial) and these will continue to have an adverse impact on global food production (FAO, 2018). Improvement across the entire smallholder linked value chain is needed to ensure a sustainable and equitable food system for all.[27]

Income derived from the sale of agricultural products is often the only source for smallholder farmers, which means their household incomes are tightly linked to agricultural productivity. Uncertainties in, and the inability to adapt to factors that influence productivity are the primary reasons smallholder farmers are at the bottom of the economic pyramid. Based on the scale at which they pose problems, risks in smallholder agricultural systems can be broadly classified into market prices, i.e. change in input and sale prices; production, which can change based on either biotic (pests/diseases) and/or climatic extremities (drought, flooding); financial issues, such as unexpected changes in interest rates or changes in non-farm incomes; and institutional and legal shocks, such as sudden changes in land use regulations and policies (Cervantes-Godoy *et al.*, 2013). In a recent meta-analysis and a global review of literature on farmers' perceptions of factors that cause productivity risks, approximately 55 percent of farmers pointed to weather-related risks, followed by pest and disease shocks

[27] See https://ciat.cgiar.org/about/strategy/sustainable-food-systems/ for food system definition.

(biosecurity risks), and human risks such as loss of family or farm labour (Duong *et al.*, 2019) Additionally, it is also becoming evident that productivity risks are complex and, more often than not, interrelated.

The opportunity of digital transformation for smallholder farmers and their value chain

Digital transformation is widespread in society. Advancements in ICTs are also benefiting smallholder farmers. Digital connectivity has improved and, in several instances, has out-competed traditional physical connectivity channels. There still exist numerous geographically remote locations in which traditional agricultural extension services have infrequent reach, but smallholder farmers in such remote areas already possess at least a basic level of digital connectivity through mobile phones.

A significant proportion of risks faced by smallholder farmers is often ascribed to the absence of timely and transparent access to actionable information, ranging from type/price of inputs to in-season advisories and post-season sales (Jellema *et al.*, 2015). The formation of information societies as a result of the digital revolution has the potential to solve information-related challenges that smallholder farmers currently face (Jellema *et al.,* 2015; Rao, 2003). Specifically, increased access to mobile phones has become central to this digital transformation in smallholder communities. This access has opened up new solutions and innovations that can address risks and pain-points across the smallholder-linked agricultural value chain.

Digital solutions can be classified on a continuum from simple to complex, based on data capture, storage, and analytics and resulting services.

Simple and scalable

Digital solutions on the simpler end of the continuum are generally characterized by one-way communication; that is, information is only pushed/delivered but is not actively collected, in order to develop improved or new services. These solutions also tend to incorporate a basic level of data analytics. Such solutions offer the advantage of adoption at scale and are also simple for the end user. They can be used on mobile phones with both simple functionalities (feature phones) and complex functionalities (smartphones).

These solutions aptly represent the foundation of information societies as they are able to connect stakeholders, and facilitate the flow of information within and across the agricultural value chain as well as among stakeholders. This connection of stakeholders and information flow leads to the development of new services that benefit not only the smallholder farmers, but also the rest of the stakeholders in the value chain. Timely and affordable access to tractors, for example, is often a pain-point for numerous smallholder farmers across Africa and Asia. Services such as Hello Tractor[28] and TunYat[29] simplify this by using a crowdsourcing

[28] https://www.hellotractor.com/home
[29] http://www.tunyat.com/

approach and a mobile application that helps connect tractor owners with those needing a tractor. Services such as MasAgro Movil[30] provide actionable in-season information and decision support by providing pest, climate, and price alerts directly to smallholder farmers via their mobile phones. Such essential information would in the absence of digital solutions come from in-person visits of extension agents or input suppliers and therefore would not be as timely, and thus as useful for the farmers. Services such as WeFarm,[31] RegoPantes,[32] and Farm Citizens[33] facilitate peer-to-peer networking between smallholder farmers in a format similar to Facebook, thereby enabling the flow of knowledge, experience and information between farmers, and removing barriers related to information and knowledge access, which has been an issue for smallholder farmers, especially in geographically remote locations.

Connection of stakeholders and information flow across the value chain does lead to the development of new services, which is evident from services such as WeFarm, mFarm,[34] RegoPantes and Kalgudi.[35] These have leveraged the advantages offered by digital connectivity to add new functions to their existing services, for example digital market places that can be used by farmers to sell their produce directly to the prospective buyers without the need for middlemen such as distributors and traders. The elimination of middlemen improves transparency, increases efficiency in the value chain, and provides improved chances for farmers to obtain fair prices for their products.

Complex solutions and their niche

A review of trends in the digital solutions ecosystem for smallholder farmers reveals that numerous solutions are attaining large-scale adoption. This has been primarily driven by the evolution and growing complexity of these solutions, some of which are bundling or providing additional features that address multiple pain-points for stakeholders along the value chain. In addition, the solutions are also becoming more complex. They are being designed to collect specific data from smallholder farmers (two-way communication) and are combining this with various other types of data (for example, seasonal forecasts and market prices). These data can then be processed using advanced predictive analytics to develop newer and more intelligent services for smallholder farmers.

Processes that result in a farmer's decision to apply fertilizers, both in terms of timing and volume, are complex. So are decision process related to the timing of irrigation, disease detection and mitigation. In all of these cases the farmer needs to consider factors such as weather, product price, traditional knowledge, as well as other factors. Access to and the availability of diverse data and information sources, as well as improvements in analytics, have led to the development of digital solutions that can automate such complex processes and support farmers in their decision making. Precision Agriculture for Development,[36] for example, uses data acquired from multiple sources, such as individual farm level management practices crowdsourced by farmers using their mobile phones, weather data and forecasts,

[30] https://movil.masagro.org/es/
[31] https://wefarm.co/
[32] https://www.regopantes.com/
[33] https://play.google.com/store/apps/details?id=com.pureforceagri&hl=en
[34] https://www.mfarm.co.ke/
[35] https://kalgudi.com/index.html
[36] http://precisionag.org/

as well as data obtained from remote sensing devices, and deploys machine learning/artificial intelligence to mine and provide location-specific agrometeorological advisories for rural smallholders across India, Kenya, Pakistan, and Rwanda. Unlike large scale farmers in the developed countries, farm level management in emerging economies varies between farms; however with advanced analytics, it is now possible to not only learn and better understand relationships between management practices, climate and yields in the past, but also it is possible to predict, and recommend the best set of practices to employ given a particular forecast. In the state of Gujarat in India, close to 1 200 cotton producing farmers signed up for the Precision Agriculture for Development service in 2015 and followed its advice, which resulted in an 8.6 percent increase in cotton yields and provided each farmer with an income increase of roughly USD 100 for that year.

Access to finance is essential, but is a challenge for numerous smallholder farmers across the globe. A combination of lack of reach, physical access, and risk information makes it difficult for the local smallholders to gain access to the capital needed to purchase basic inputs such as fertilizers, which could potentially help them obtain higher yields and incomes. Agribuddy[37] uses a mixed approach of physical and digital interfaces. It recruits rural youth to act as "buddies," who are responsible for persuading farmers to enroll in the service, and for collecting and monitoring farm characteristics and productivity indices from the enrolled farmers using AgriBuddy's mobile application. Data collected on the platform is then used to develop profiles for smallholders, which are shared with a financial services provider. Based on the profiles, the financial services provider can decide whether or not to extend loans to the farmers. In addition to developing risk profiles, AgriBuddy leverages the profile data, deploys advanced analytics to predict yields and forecasts, and further strengthens the farmers' profiles with respect to on-farm productivity and crop management practices, and therefore provides more information about the farmers with improved insights, to the financial services providers. AgriBuddy – with its business model of promoting rural employment, in combination with developing targeted services for smallholder farmers – is on track to facilitate credit access for close to two million smallholder farmers in Asia by 2020, and generate additional employment of 50 000 young, rural residents to act as buddies by that year.

Access to export markets and commodity crop value chains positively affects the incomes of smallholder farmers, as it provides an opportunity for both sustained and improved incomes. Exporting, however, requires suppliers to comply rigorously with globally acceptable sustainability standards such as GlobalGAP.[38] A smallholder farmer needs to undergo rigorous evaluation of his/her management practices every season by certifiers from GlobalGAP in order to obtain certification. Verifik8[39] employs locally-recruited staff to enroll farmers, who collect farm characteristics and crop management information in near-real time, using a mobile application. The data is analyzed to develop key metrics, which enables a farmer to obtain the necessary certification, thus, making them compliant with the export market. Sustainable sourcing is a key focus for Thai Union, a company that is involved in the production and sale of seafood. However, smallholder shrimp farmers need to obtain necessary social and environmental certification in order to sell to Thai Union. Verifik8 bridges this gap between

[37] https://www.agribuddy.com/
[38] https://www.globalgap.org/uk_en/
[39] https://www.verifik8.com/

Thai Union's certification need and shrimp farmers' need for export markets by developing a solution that collects socio-environmental data from shrimp farmers (e.g. social indicators such as whether or not a farmer uses child labour and environmental indicators such as a farm's energy use), which is further verified using voice call surveys of shrimp farmers. Validated data is then converted into metrics and certification that can be used by the shrimp farmer to sell the produce to Thai Union.

Land ownership is essential for smallholder farmers as this enables them to access credit from banks, and/or sell produce to value chains maintained by multinational companies; however, land registration data in the developing world is prone to fudging and tampering. A single piece of farmland can have multiple owners, which prevents the actual owner from gaining benefits such as access to loans and credit. Bitland[40] works together with farmers and banks in Ghana to digitize farmer land records using blockchain systems. A blockchain system ensures that the land registration information is tamper-proof. Data collection agents at Bitland collect land registration information from the national land registry, produce a report, and secure the data by storing it on their blockchain system. By showing the report to the bank, the registered farmer becomes eligible to access financial services.

Pest and disease attacks are a major factor limiting smallholder farmer productivity and income. At its core, the problem is in the time lag between actual disease occurrence, identification, and mitigation. As most of the damage is already done once the disease is correctly identified, there is a need to detect and confirm the disease quickly, as well as suggest suitable remedies. Mobile phone-based disease diagnostic solutions, such as Nuru[41] and Plantix[42] are designed to address these issues and reduce this time lag. Whereas Nuru provides a near real-time diagnosis of a limited set of diseases, Plantix provides a diagnosis of a larger set of crop diseases. In addition, Plantix is also designed to detect abiotic stresses such as drought stress and also provides treatment recommendations based on the disease identification. Whereas Nuru and Plantix use images of diseased plants captured through farmers' mobile phones, Farm.Ink[43] harvests unstructured data gathered from chat exchanges and pictures in dedicated livestock farmers' Facebook groups, through a chatbot that automatically analyses the chat and picture exchanges, and provides disease diagnosis, suitable remedies, and other information to the group in near-real time.

To learn more about a host of other different digital solutions and their providers, the reader is directed to review GSMA's mAGRI development tracker.[44]

The drivers and impacts of data flow – big data in action

This observed evolution of digital solutions and their ability to tackle the complex, multiple challenges smallholder farmers face can be ascribed to a number of key technological advancements.

---

[40] http://landing.bitland.world/
[41] https://www.iita.org/tag/nuru/
[42] https://plantix.net/en
[43] https://farm.ink/
[44] https://www.gsma.com/mobilefordevelopment/m4d-tracker/magri-deployment-tracker/

(a) The bidirectional movement of data and information between smallholder farmers and service providers at speed has been made possible through advances such as the evolution of networks from 2G to 3G and upwards, for example, which enables the faster transfer of greater data volumes across mobile networks (Gawas, 2015).

(b) Concurrent with advancements in network technologies, the cost to obtain mobile phone technology has decreased such that many more people, irrespective of their economic or social status, can now afford more powerful smart phones (The Economist, 2014) that enable them to complete many more complex tasks.

(c) The "Open Data" movement and improvements in satellite technologies also deserves a special mention, as these have in combination facilitated access to remote sensing and satellite data sources that have become an integral part of the bouquet of data sources in action (Harris and Baumann, 2015).

(d) The hype and buzz around artificial intelligence and machine learning is fading as new use cases associated with solving challenges related to basic needs and access are emerging. The focus has moved towards grounding artificial intelligence through improved statistical rigour in interpretation, algorithm advances in deep learning, and new reinforcement learning approaches, all of which enable solutions with greater precision and accuracy to be developed.

Although advanced technology and concepts such as precision farming for smallholder farmers using drones and sensors are still in their infancy, most of these technological advances are indeed pushing the boundaries. Their use by smallholder farmers is gradually becoming mainstream.

Farming systems – more specifically, smallholder systems – are complex (Giller *et al.*, 2011). Before the advent of big data, approaches to address productivity risks relied on system simplification (sometimes over-simplification) and solutions were devised via reductive reasoning, using limited data/information. The deluge of data collected from smallholder farms, from diverse streams of data (e.g. mobile app collected, drone collected, sensor collected) on the same farm, is contributing towards the generation of digital profiles of smallholder farmers, a concept that has been thoroughly reviewed by USAID and the Grameen Foundation (Gray *et al*., 2018). These digital profiles are now being leveraged in multiple, creative ways to address pressing issues, some of which limit the efficiency and productivity of smallholder farmers and their value chains.

FAIR data warehouses and data lakes – the next wave

Data and digital technologies have made processes across diverse fields (e.g. medicine, industrial production) more intelligent and efficient. The success of exemplar solutions mentioned in this report and many others does provide sufficient evidence that data and digital tools, as in other fields, have made the processes associated with smallholder farmers and their value chains intelligent and efficient, and have made a positive impact on smallholder farmer incomes and livelihoods. At its core, this positive impact has been made possible because of the fact that data have been able to flow seamlessly across multiple sources, thus enabling integrated and advanced analysis and the development of smarter and more precise solutions.

The “dark horse” that has not been sufficiently recognized, but which has been an integral part of the development of improved simple and complex digital services for smallholders, is the data infrastructure. A data infrastructure is a structure consisting of hardware and software processes that enables data capture, storage, flow, and analytics. These data infrastructures are typically referred to as data lakes or data warehouses. The data contained within these data lakes and data warehouses are essential for the development of digital profiles of smallholder farmers and their innovative use is further driving innovations across digital solutions.

Although digital profiles, data warehouses and lakes are starting to emerge in the smallholder data ecosystem, they are not unified and cannot be leveraged in unison, as different service providers use varying combinations of data sources and algorithms to create smallholders’ digital profiles. To put it simply, data warehouses and data lakes from different service providers currently do not and cannot talk to each other!

The next wave of digital solution transformations will depend on the creation of findable, accessible, interoperable, and reusable (FAIR) data lakes and warehouses by stakeholders in the value chain. The FAIR concept was previously associated with data, but, with the emergence of big data infrastructures, and the evident improvement in services as a result of leveraging multiple sources of data, FAIR data lakes and data warehouses are also needed.

Specific stakeholders in the value chain, such as large agribusinesses that provide physical and digital services at multiple points in the value chain – those that sell inputs, and may also provide credit access and/or buy agricultural produce, etc. – are better positioned to develop new, intelligent, actionable data-driven services for smallholders, as they contribute to, manage, and use a significant proportion of the existing data lakes and/or data warehouses currently present in the smallholder farmer data ecosystem. The success of these remains dependent on their ability to consolidate the internal data infrastructures.

The data ecosystem around smallholder value chains is vast and varied. Therefore, there is, and will always be, an opportunity for several small- and medium-sized service providers to leverage the multiple data sources available to apply advanced analytics and deliver smart solutions. Independent of the size of the service providers and their contribution to the data ecosystem of smallholders, there will be a need for institutional, technical transformations, followed by business process innovations, to make disparate data lakes and data warehouses work for smallholder agricultural systems.

More specifically, data management systems and procedures across multiple services need to be transformed, but without sacrificing the competitive advantages of individual services. Interoperability across institutional data policies, as well as data and metadata storage systems, and/or the creation of reproducible and open algorithms, will be the key to fueling the next wave of big data use for smallholder farmer value chains.

Those data management frameworks that guarantee FAIR access to data lakes and/or warehouses, without sacrificing data integrity, will be essential for enabling the next wave of transformations. Blockchains, for example, have the potential to facilitate data lake/data warehouse interoperability without sacrificing data integrity. Various data types, which are assets, can be securely transferred in a decentralized manner using this technology.

A wider data ecosystem level transformation for enabling data lake/warehouse interoperability is being facilitated by platforms such as the Open Project Algorithms (OPAL) (Letouzé and Pentland, 2018). This socio-technological innovation was originally designed to enable access to private sector data gathered for the public good, in a privacy preserving, predictable, participatory, scalable, and sustainable manner. The OPAL framework was originally piloted with telecom operators Orange Sonatel and Telefónica Colombia. It was used to request specific indicators (e.g. population densities). The request essentially is a pre-developed algorithm, which when sent to the telecom operator was implemented inside the telecom operator firewall. The results are made available to the requester through a different interface. The algorithms are open for public use and this gave the project its name. In addition to this technology layer, there is also a social layer, wherein every step – from sending the request, to implementation of the algorithm, and sharing of results – needs to go through an approval process by a local advisory committee on ethics and development. This committee provides oversight and guidance, as well as ensures that the entire process follows data ethics principles. Although OPAL was designed to leverage private sector data and not truly a data lake or a warehouse for public good, it can potentially be employed to leverage privately-held data lakes in order to develop better solutions for smallholder farmers. Several existing, secure data management frameworks (e.g. Globus) can also be adapted to render data lakes, and/or data warehouses FAIR.

Data products and solutions from, for example, various agrifood systems research institutions, which are also key components of the smallholder farmer data ecosystem, are also undergoing transformation. These institutions are making their data repositories FAIR. CGIAR, for instance, is developing its data and information discovery tool called Global Agricultural Research Data Innovation & Acceleration Network (GARDIAN).[45] The tool's use of data discovery algorithms (e.g. graph theory, ontologization) can potentially facilitate the improved alignment of research data and solutions with data associated with the existing smallholder data ecosystem. This will further catalyze the use of data and digital tools for developing solutions that can positively impact smallholder farmer livelihoods.

The ability of stakeholders to collate data across multiple data warehouses and/or data lakes will undoubtedly enable the next wave of evolution in digital solutions. Agribusinesses with large-scale operations, which are already in the business of providing multiple physical services across the value chain – such as providing a variety of inputs (e.g. seeds, fertilizers and credit) to farmers and that also buy produce from smallholder farmers, at scale – are ideally placed to be the “first-movers”, and will be able to showcase the advantages of enabling disparate data warehouses to interact.

The data ecosystem of smallholder farmers consists of several digital solution providers, whose solutions, on the basis of their data processes, can be placed on a continuum ranging from simple to complex. Regardless of the type of the solution, each of these has its advantages. Whereas relatively simple solutions have a greater potential to be adopted at scale, the more complex applications address key, complex, niche issues, but their adoption potential is limited. More importantly, each of these solution types still has a market potential of roughly 570 million smallholder farmers to whom it can cater. The complex solutions have

[45] http://gardian.bigdata.cgiar.org/#!/

also shown the power of data fusion and advanced analytics to address complex issues related to farming. The digitization of processes across smallholder farmer value chains and the generation of digital profiles of farmers have been shown to have immense potential in the development of solutions that improve agricultural productivity and smallholder farmer incomes. In the current phase of technology transformation, we can observe the presence of a diverse range of solution providers, all holding immensely powerful digital profiles. Moreover, data integration and advanced analytics have shown the potential to develop innovative solutions. Therefore, the next wave of digital solution transformation will involve the development of big data frameworks that will facilitate interaction between data lakes and data warehouses that will further lead to the creation of many more innovative and intelligent services designed to help make smallholder farmers more productive and their value chains more efficient and transparent.

References

CCAFS. 2014. *Cracking patterns in big data saves Colombian rice farmers' huge losses* [online] Wageningen. [Cited 19 March 2018] https://ccafs.cgiar.org/research/results/cracking-patterns-big-data-saves-colombian-rice-farmers%E2%80%99-huge-losses#.XMMvSegzbIV

Cervantes-Godoy, D., Kimura, S. & Antón, J. 2013. Smallholder risk management in developing countries. *OECD Food, Agriculture and Fisheries Papers*, No. 61 [online] Paris. [Cited 23 January 2019] http://dx.doi.org/10.1787/5k452k28wljl-en

Duong, T., Brewer, T., Luck, J. & Zander, K. 2019. A global review of farmers' perceptions of agricultural risks and risk management strategies. *Agriculture*, *9*(1): 10.

FAO. 2018. *The State of food and agriculture 2018. Migration, agriculture and rural development*. Rome. 174 pp. (Also available at http://www.fao.org/3/I9549EN/i9549en.pdf).

FAO. 2014. *The state of food and agriculture 2014. Innovation in family farming*. Rome. 139 pp. (Also available at http://www.fao.org/3/a-i4040e.pdf).

Gawas, A.U. 2015. An overview on evolution of mobile wireless communication networks: 1G–6G. *International Journal on Recent and Innovation Trends in Computing and Communication*, *3*(5): 3130–3133.

Giller, K.E., Tittonell, P., Rufino, M.C., Van Wijk, M.T., Zingore, S., Mapfumo, P., Adjei-Nsiah, S., Herrero, M., Chikowo, R., Corbeels, M. & Rowe, E.C. 2011. Communicating complexity: integrated assessment of trade-offs concerning soil fertility management within African farming systems to support innovation and development. *Agricultural Systems*, 104(2): 191–203.

Gray, B. Babcock, L., Tobias, L., McCord, M., Herrera, A, Osei, C. & Cadavid, R. 2018. *Digital farmer profiles: reimagining smallholder agriculture*. United States of America, USAID. (Also available at https://www.usaid.gov/sites/default/files/documents/15396/Data_Driven_Agriculture_Farmer_Profile.pdf).

Harris, R. & Baumann, I. 2015. Open data policies and satellite Earth observation. *Space Policy*, 32: 44–53.

Jellema A., Meijninger W. & Addison C. 2015. Open data and smallholder food and nutritional security. *CTA Working Paper* 15/01. Wageningen: CTA.

Lesiv, M., Laso Bayas, J.C., See, L., Duerauer, M., Dahlia, D., Durando, N., Hazarika, R., Kumar Sahariah, P., Vakolyuk, M.Y., Blyshchyk, V. & Bilous, A. 2019. Estimating the global distribution of field size using crowdsourcing. *Global Change Biology*, 25(1): 174–186.

Letouzé, E. & Pentland, A. 2018. Towards a human artificial intelligence for human development. *ITU Journal: ICT Discoveries*, 1(2).

Lowder, S.K., Skoet, J. & Raney, T. 2016. The number, size, and distribution of farms, smallholder farms, and family farms worldwide. *World Development*, 87: 16–29.

Rao, M. 2003. *The nature of the information society: a developing world perspective* [online] Geneva. [Cited 07 March 2019] http://www.itu.int/osg/spu/visions/papers/developingpaper.pdf

Ricciardi, V., Ramankutty, N., Mehrabi, Z., Jarvis, L. & Chookolingo, B. 2018.

How much of the world's food do smallholders produce? *Global Food Security*, 17: 64–72.

The Economist. 2014. The rise of the cheap smartphone [online] London. [Cited 11 March 2019] https://www.economist.com/business/2014/04/05/the-rise-of-the-cheap-smartphone

World Bank. 2016. A year in the lives of smallholder farmers. 25 February [online] Washington, DC. [Cited 28 March 2019] http://www.worldbank.org/en/news/feature/2016/02/25/a-year-in-the-lives-of-smallholder-farming-families

Contact

Dr Dharani Dhar Burra (d.burra@cgiar.org)
Article edited by Linda Dailey Paulson

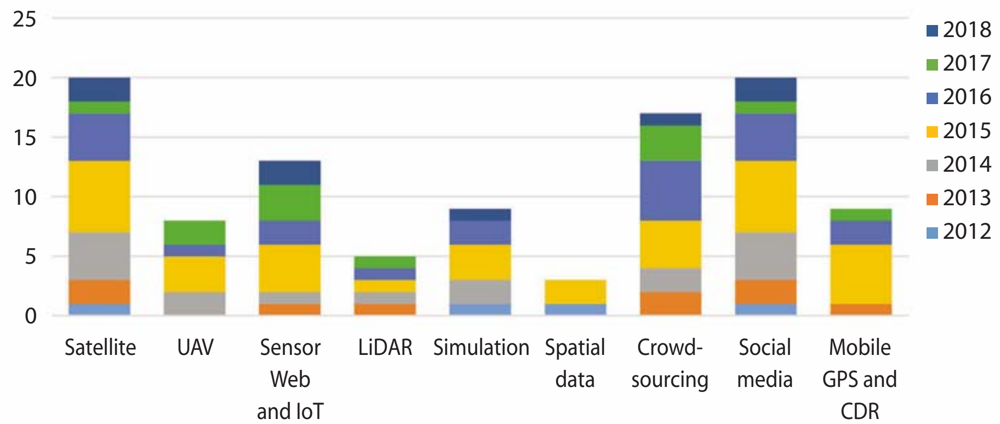
Disclaimer: The mention of specific companies or products of manufacturers, whether or not these have been patented, does not imply that these have been endorsed or recommended by CGIAR Big Data Platform for Agriculture.

Big data ecosystem for disaster resilience

Big data has evolved at an unbelievably fast pace. In the specific context of disaster resilience, big data can help in all four phases of disaster management: prevention, preparedness, response, and recovery. Two major sources of big data, namely dedicated sensor networks (e.g. earthquake detection using seismometers) and multi-purpose sensor networks (e.g. social media such as Twitter using smartphones), have demonstrated their usefulness in disasters such as the Tohoku Earthquake (Anderson *et al.*, 2013). However, two of the major big data challenges continue to exist in several countries: variety (integration of many data sources including dedicated sensors and multi-purpose sensors), and veracity (filtering of big noise in big data to achieve high quality information).

The data from the emerging technologies including satellite imagery, aerial imagery and videos from unmanned aerial vehicles (UAVs), sensor web and Internet of things (IoT), airborne and terrestrial Light Detection and Ranging (LiDAR), simulation, spatial data, crowdsourcing, social media, and mobile GPS and Call Data Records (CDR) form the core of big data for disaster resilience (Yu *et al.*, 2018). Figure 10 illustrates that satellite imagery, crowdsourcing, and social media are becoming more popular sources of data for disaster risk management.

Big data has opened up promising approaches to disaster resilience. Mobile phone data, for example, can provide an incredibly detailed view of population behaviour and movement in areas that were previously observed infrequently and indirectly. Social networks such as Twitter



Source: Yu *et al.* (2018)

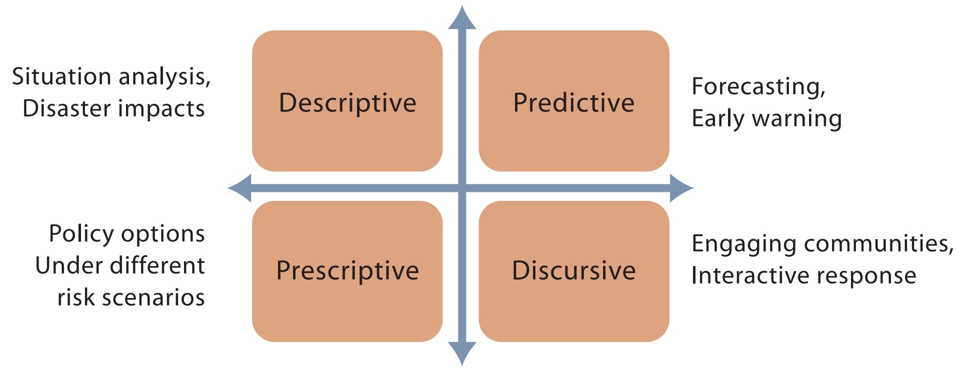
Figure 10: Increasing use of satellite imagery, crowdsourcing, and social media for disaster management

and Facebook are already improving the ability of humanitarian and disaster risk reduction organizations to monitor and respond to hazards. Furthermore, opportunities are increasing as mobile phone penetration and access to the Internet, for example, are increasing significantly in high disaster risk, developing countries.

How does big data strengthen resilience?

Disaster resilience calls for a wide spectrum of data driven activities and actions. For example, big data fills in the critical data gaps to operationalize the emerging trends in multihazard early-warning systems to provide impact-based, risk-informed, people-centered and end-to-end early-warning services at different scales: regional, subregional, national, local, and community level. It also enables the transition from early warning to early action such as in the cases of forecast-based financing, forecast-based social protection and risk prevention.

Big data enables descriptive, predictive, prescriptive and discursive analytics that addresses gaps in information flows in pre-disaster, response and post-disaster situations (Figure 11).



Source: Data-pop alliance synthesis report (2015)

Figure 11: Big data analytics functions for disaster resilience

Big data solutions for rapid damage assessment

As part of the World Bank's response to the Sulawesi earthquake and tsunami, a rapid assessment of the damage affected areas in Central Sulawesi was conducted using the Global Rapid post-disaster Damage Estimation (GRADE) methodology. It was the first report to produce sector-based preliminary economic loss estimates, based on scientific, economic and engineering data and analysis that informed disaster recovery and reconstruction processes. The assessment was based on an open loss modeling approach developed by the World Bank's Disaster-Resilience Analytics and Solutions team, that includes:

- analysis of satellite imagery and other ground-collected data (Indonesian National Board for Disaster Management (Badan Nasional Penanggulangan Bencana), Ministry of Education and Culture (Kementerian Pendidikan dan Kebudayaan), Ministry of Public

Works and Housing (Kementerian Pekerjaan Umum dan Perumahan Rakyat), the ASEAN Coordinating Centre for Humanitarian Assistance on Disaster Management, media etc.);

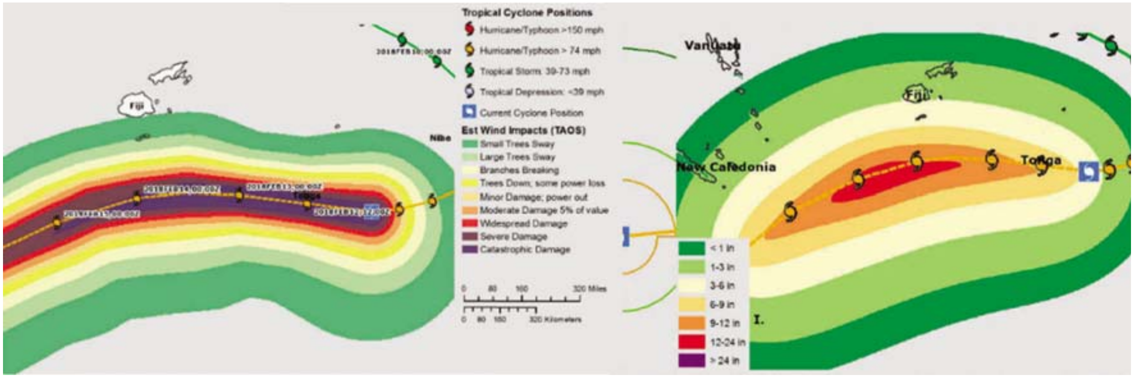
- remote-sensing imagery (UNOSAT, COPERNICUS, DigitalGlobe, Google, Humanitarian Open Street Map team, MapAction);
- information coming out of early assessments, as well as social media data for results calibration; and
- spatial characteristics developed for tsunami events that included inundation extent and ground deformation analysis that were produced based on a remote sensing damage assessment.

Source: International Disaster Charter (2018)

Figure 12: Pre-and post disaster satellite images of Sulawesi's earthquake and tsunami affected area for damage assessment.

The main benefit was the speed at which the damage estimation was produced. Within 10 to 14 days of an event, stakeholders were able to access loss estimates and spatial distribution of damages. The assessment of damage was made using pre-and-post disaster scenarios – captured by time series remote sensing data in conjunction with ground-based information including drone and social media reports (Figure 12). The World Bank was able to make rapid estimates – total economic damages of USD 500 million, USD 180 million for the housing sector, USD 185 million for commercial/industrial buildings, and USD 165 million for infrastructure – to programme its support for Sulawesi recovery and reconstruction efforts. The World Bank announced funding of up to USD 1 billion for the Government of Indonesia to supplement relief and reconstruction efforts in the disaster-affected areas of Lombok and Sulawesi, and to bolster long-term resilience (World Bank, 2018).

Another example is the Cyclone Gita impact-based forecasting and damage assessment. Tropical Cyclone Gita hit several countries in the Pacific with Samoa first to be hit, followed by Niue, Tonga, and Fiji from 10 February to 13 February 2018 (Reliefweb, 2018). The cyclone was predicted well in advance and accordingly preparedness measures were put in place



Estimated wind impacts

Estimated rainfall

Source: Pacific Disaster Center, 2018

Figure 13: The impact-based forecasting of Cyclone Gita

by the government (Virma, 2018). The use of big data helped to enable the impact-based forecasting of Gita. The cyclone tracks, the wind radius of the cyclone, rainfall estimation and wind impacts played a key role in preparing the countermeasures (Figure 13).

Furthermore, Tonga's post-disaster needs assessment was carried out using unmanned aerial vehicles (drones). The advantage of a drone's aerial images compared to satellite imagery was its higher resolution (below one metre), which was an important requirement for small area damage estimation (Virma, 2018). Drone images were particularly useful to capture damaged buildings and infrastructure, land cover assessment and economic loss. It also helped in rapid mapping and consequently accelerated the reconstruction and recovery process.

Forecasting and early warning

Big data fills in the critical data gaps that exist in operationalizing flood forecasting and early warning. The recent advances in climate modeling such as ensemble prediction system (EPS) indicate a promising trend in flood forecasting with a long lead time. The EPS is often better than single (deterministic) forecasts as it shows the possibility of severe rainfall in case the single forecast failed to capture it. The EPS also estimates forecast uncertainty from the width of the ensemble spread (Figure 14). The EPS helps substantially in flood forecasting and early warning in transboundary river basins, often constrained by lack of access to and availability of hydrologic data. This approach indicates the benefit in incorporating rainfall predictions from multiple weather centres, as well as rainfall and river observations from multiple platforms and institutions. For some stations, skillful forecast lead times are as long as 16 days (Lnu *et al.*, 2017).

The experience of EPS for 2018 flood forecasting in Sri Lanka was a mixed success. Although it captured the intensity of torrential rain two days in advance, the forecast was not precise in its exact location (Figure 15) (Ushiyama, 2019). The location accuracy can be improved not only with better quality of downscaling ensembles but also with densification of the data network and putting in place an appropriate a big data ecosystem.

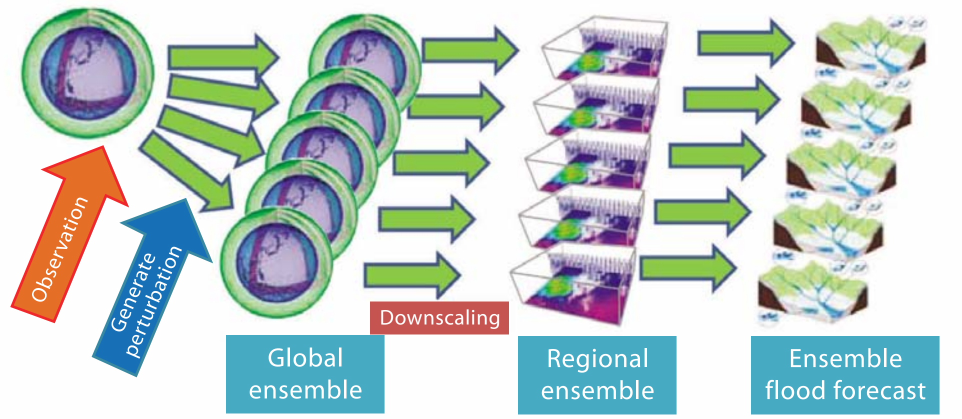


Figure 14: Ensemble prediction system: nested modeling for flood forecasting with longer lead time

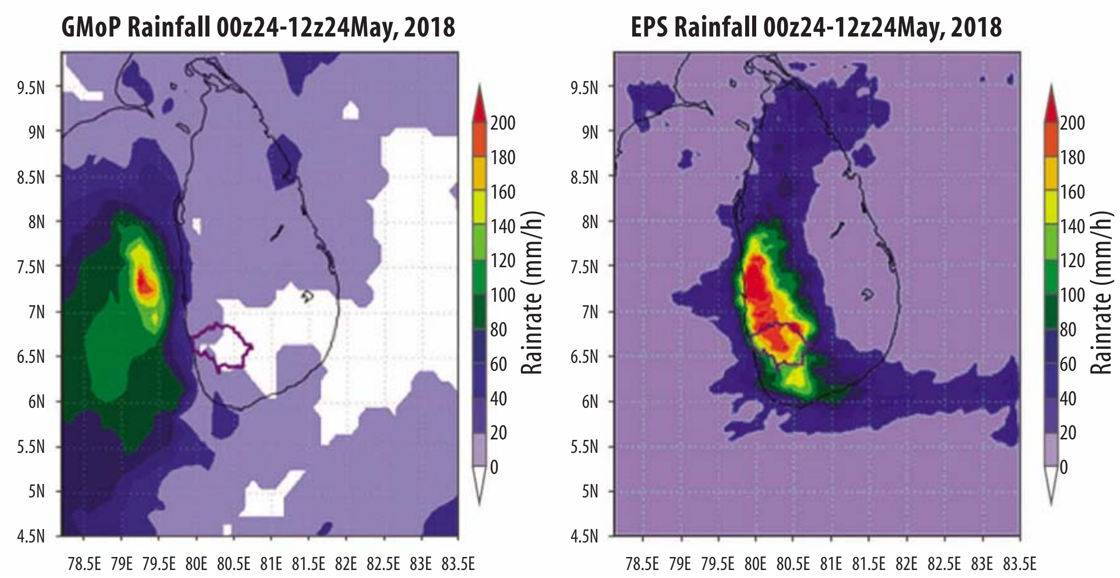


Figure 15: Challenges of ensemble prediction system for flood forecasting in Sri Lanka

The emerging trend of impact-based flood forecasting is realized with the big data applications. The big data based system approach allows end users to build impact scenarios of the forecasted event by crossing real-time data on flood hazard, exposure and vulnerability. With a web-GIS platform it is possible to aggregate data both in a temporal or spatial way and to build scenarios of risk and damage to develop impact-based flood forecasting (Rossi *et al.*, 2017).

Applications of machine learning for disaster resilience

Resilience building relies on many different data types, information sources, and types of models to be effective. Even experts can struggle to develop models that enable the understanding of the potential impacts of a hazard on the built environment and society. The advances in machine learning offer new approaches and new ways to have more accurate, efficient, and useful solutions for resilience building. A machine learning (ML) algorithm "learns" from previous data and its output is a result that adds information and insight that was not previously known. This approach enables actions to be taken on the information gathered from the data; sometimes in near real time, such as suggested web search results, and sometimes with longer-term resilience building based on multiple risk scenarios (Deparday *et al*., 2019). Machine learning is a subset of artificial intelligence (AI), but the two terms are often used interchangeably. The seamless linkage of ML with data mining and big data ecosystem (Figure 16) – from image, sound, and voice recognition features of our smartphones, for example, enables disaster managers to identify where people are at risk in a natural disaster as measured by locations and types of buildings.

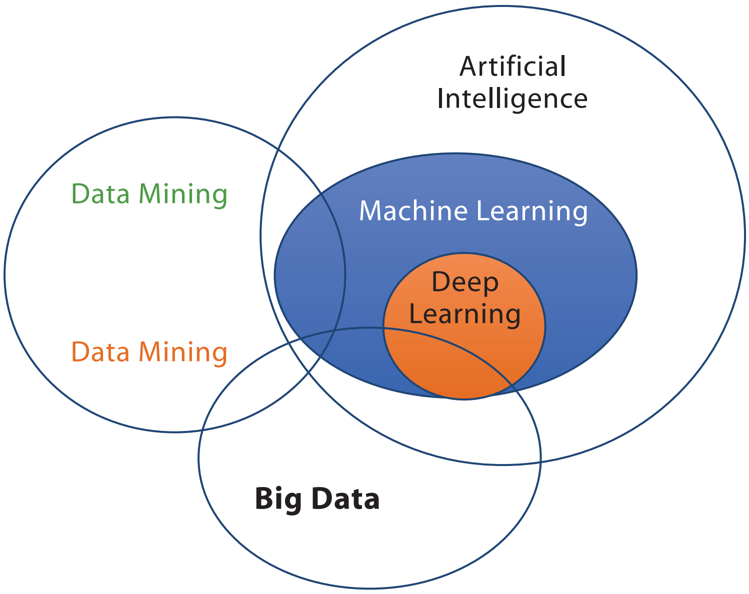
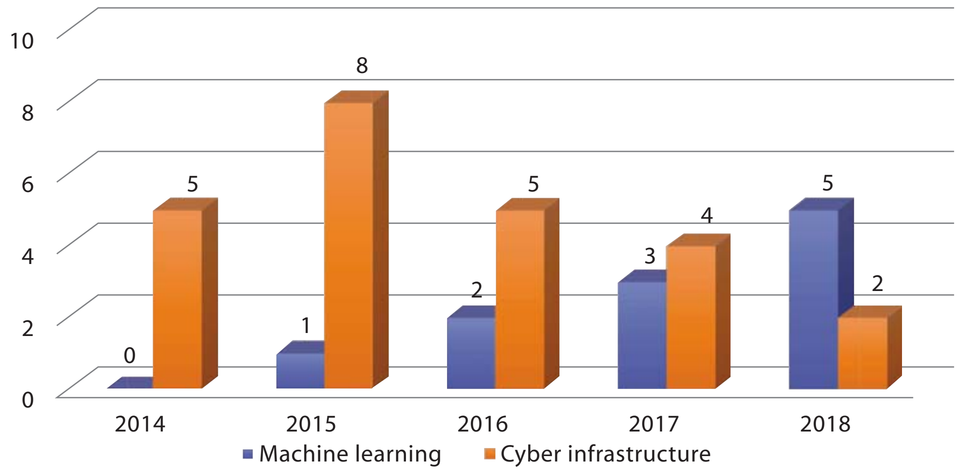


Figure 16: Machine learning – a subset of artificial intelligence and linked with data mining and big data

Managing disasters becomes more efficient if data are acquired from different sources in a higher spatial and temporal resolution. However, challenges emerge because of the constantly increasing quantum of image and video data. Emerging technological innovations including social media, location-based systems, radio frequency identification, and big data analytics are considered promising powerful tools that may help during the disaster management cycle. Processing and analyzing the heterogeneous big disaster data require efficient data collection, aggregation, information extraction, visualization, and efficient distribution. The growth of data and the need for an efficient distribution makes the development and operation of cyber infrastructure very demanding.

Machine learning has evolved to become one of the most effective methods of helping in eliminating unrelated data and speeding up all risk analytics to identify the most optimal response actions and resilience strategies. The tools of descriptive, predictive, prescriptive and discursive analytics can now leverage ML in a big way. The emerging trends clearly indicate that the ML applications in disaster resilience contexts are edging out the cyber infrastructure (Figure 17).



Source: Yu *et al*. (2018)

Figure 17: Emerging trend of machine learning, from 2014 to 2018

The key challenges in large scale operationalization of big data include: (i) challenges of dealing with the large variety of heterogeneous data from different data sources – from sensors to crowdsourcing, including time series, semi-structured and invalidated data, and textural data; also noise and misinformation; (ii) big data analytics – analytics yet to integrate reliably and accurately crowdsourced data from the disaster-affected people into the physical sensing data (e.g., satellite, UAV) and authoritative data (e.g. terrain data, census data); and (iii) cyber infrastructures for the effective integration of huge data from multiple sources for real-time decision making.

References

Anderson K., Arora, A., Aoi, S., Fujinuma, K., *et al*. 2013. Big data and disaster management, *Technical Report No. GIT-CERCS-13-09.* Georgia Institute of Technology, CERCS. A Report from the JST/NSF Joint Workshop, JST/NSF Joint Workshop Report on Big Data and Disaster Management. Editors, C. Pu & M. Kitsuregawa.

Data-pop Alliance (2015). Big data for climate change and disaster resilience: realising the benefits for developing countries. Data-pop alliance synthesis report, September 2015.

International Disasters Charter. 2018. Earthquake and tsunami in Indonesia [online]. [Cited 13 March 2019] https://disasterscharter.org/web/guest/activations/-/article/earthquake-in-indonesia-activation-587-

Lnu, S.P., Young, W., Hopson, T. M. & Avasthi, A. 2017. *Flood risk assessment and forecasting for the Ganges-Brahmaputra-Meghna River basins (English)*. Washington, DC, World Bank Group. (Also available at (http://documents.worldbank.org/curated/en/272611508255892547/Flood-risk-assessment-and-forecasting-for-the-Ganges-Brahmaputra-Meghna-River-basins).

Pacific Disaster Centre. 2018. Tropical Cyclone Gita. Warning 13, 12 December 2018 0300 UTC. (Also available at https://reliefweb.int/sites/reliefweb.int/files/resources/PDC%20-%20TC%20Gita%20Impacts%2C%20Warning%2013%2C%2012FEB18%200300%20UTC.pdf).

Reliefweb. 2018. Tropical Cyclone Gita – Feb 2018 [online] Geneva. [Cited 18 January 2019] https://reliefweb.int/disaster/tc-2018-000102-ton

Rossi, I., Corina, A. & Pecora, S. 2017. Dewetra platform initiative by the WMO Commission for Hydrology and the Italian Department of Civil Protection. A data sharing, multi-hazard forecasting and early warning system available for any WMO member [online]. Geneva. [Cited 18 January 2019]. https://www.wmo.int/pages/prog/drr/documents/mhews-ref/posters-pdfs/7.101%20-%20Rossi%20L%20et%20al%20Dewetra%20Platform%20MHEWC%202017%20poster.pdf

Ushiyama, T. 2019. Real time ensemble forecasting for flood early warning in Sri Lanka. Third Plenary Session for Platform on Water Resilience and Disaster in Sri Lanka, International Centre for Water Hazard and Risk Management (ICHARM), 20 February 2019, Colombo, Sri Lanka.

Virma, M. 2018) [online]. Malmö. [Cited 19 January 2019]. http://wpmu.mah.se/nmict182group2/2018/10/19/the-role-of-uavs-in-cyclone-gita-response-and-recovery-in-tonga/

World Bank. 2018. World Bank Press Release 14 October 2018. World Bank Announces $1 bn Assistance for Indonesia Natural Disaster Recovery and Preparedness. (Also available at http://www.worldbank.org/en/news/press-release/2018/10/14/world-bank-announces-assistance-for-indonesia-natural-disaster-recovery-and-preparedness).

Yu, M., Yang, C. & Li, Y. 2018. Big data in natural disaster management: a review. *Geosciences* 8, 165, doi:10.3390/geosciences8050165

Contact

Sanjay K Srivastava

srivastavas@un.org | www.unescap.org

Chief, Disaster Risk Reduction

ICT and Disaster Risk Reduction Division (IDD)

United Nations Economic and Social Commission for Asia and the Pacific

Olam Farmer Information System (OFIS): improving smallholder productivity and livelihoods

Introduction

The global supply of cash crops such as cocoa, coffee, cashew and cotton is predominantly sourced from thousands of individual, smallholder farmers, living at subsistence levels in some of the world's most remote regions. This creates significant challenges for farmers and the farming communities they support to connect to the global economy.

Although sustainability programmes from global agribusinesses have provided vital group training and other support to improve yields and incomes, until now these smallholder farmers have not have access to truly personalized advice.

Previously, accessing detailed information about these farmers and their farms had been a struggle with field staff having to collect information painstakingly using pen and paper, a highly laborious process, significantly limiting use and scalability.

Olam Farmer Information System (OFIS) technology allows field staff to survey and record, on the spot, thousands of farms, the surrounding landscape, as well as the farmer's social circumstances. This means that field staff can get better advice on interventions and compare progress on these, as well as identify hotspots for risks such as deforestation and child labour. Additionally, OFIS gives the farmers and Olam greater insights to enable them to tackle issues ranging from poor yields to climate change and child labour. And "geo-tagging" each farm establishes the first leg of traceability, providing assurance to end users on product provenance.

OFIS was launched in 2014, initially in Côte d'Ivoire as an initiative to better understand the farm landscape of Olam's cocoa suppliers. Since then, the platform has been rolled out across eight product categories, including coffee, cashew, cotton and rice, in 27 countries. By the end of 2018, 248 850 farmers had been registered on the platform. OFIS is implemented across Olam's global smallholder network in 27 countries, including Côte d'Ivoire, Brazil, Ghana, Indonesia, Mexico, Turkey and Viet Nam.

Connecting farmers to the global economy through OFIS delivers mutual benefits for smallholders and customers. Farm data and analysis enables farmers to receive more tailored support to help them improve their yields and quality, which in turn is rewarded with quality and sustainability-based premiums.

Olam's sustainability partners and customers get improved traceability and transparency with direct access to farmer and origination information. Insight derived from intervention reduces supply chain risk and improves funding efficiency.

Methodology

Data – including farm size, location, age of (tree) stock, economic, social and health infrastructure, and eco-support systems – is collected at farm-gate level on a hand-held device by Olam's trained field staff. OFIS is capable of capturing the data offline, which is important given the often remote locations of smallholder farming communities and it also enables better face-to-face interaction of the field staff with the farmers.

As a standard protocol, Olam field staff brief the farmers on how the data will potentially be used and assure them that any personal data would be available only to Olam staff and selected Olam customers under an appropriate Terms of Usage agreement and otherwise in strict confidence. After the briefing, the farmer either consents to provide relevant information or opts out from the exercise. The OFIS database is owned by Olam International Limited in Singapore and all personal data-related processes and protocols are in accordance with Singapore laws or practice.

GPS is widely used to geotag farmers' locations, plotting farms and mapping key social infrastructure points in the villages. The data from this mapping and the farmer surveys are fed via an Android OS application into Olam's database and then visualized via an online maps interface and analysis graphing tool to generate reports.

The analysis serves to highlight deficiencies in farmer or community resources, as well as farm practices, and allows Olam to manage risk and tailor action and support.

As part of Olam's sustainability interventions under the Olam Livelihood Charter, OFIS can also give personalized farm development plans to each farmer with advice on how to make the most of each plot and crop.

Impact

Through OFIS, farmers understand their farms better, take action themselves to improve the yield and quality of their crop, and ultimately having more power over their futures, which have been precarious to date.

From the OFIS data, 60 828 individual Farm Development Plans (FDPs) have been generated for cocoa smallholders – something these farmers have never had before. These personalized plans provide support that is absolutely tailored to each farm, such as the precise amount of agricultural inputs that is required, the number of shade trees to plant to protect the crop, or how to prune correctly. If the yield does not appear as expected, then Olam compares with other farmers' performances in the region and identifies what the barriers are. Having a plan empowers the farmer, and progress against the recommendations is tracked over time and adapted where necessary.

In Ghana, the tailored advice provided in these FDPs has helped some cocoa farmers more than triple their output over the last two years and reduce reliance on pesticides. The technology puts them in direct contact with advisers, so when, for example, there is an outbreak of disease, they understand that resorting to chemicals is not the only answer.

By increasing farmer productivity, the data are also helping to reduce deforestation risks as farmers are dissuaded from expanding the farm area. Equally, OFIS helps Olam to work with farmers by monitoring farm boundaries and suggesting how farmers might diversify incomes with other crops. To date, 187 230 farms have been mapped globally. Forestland is very rich and fertile and preventing forest encroachment is essential for holding back climate change, which is already impacting the production of commodities such as cocoa and coffee.

Meanwhile, rural communities are supported with improved awareness of social needs. By mapping information on local infrastructure and key social metrics for instance, Olam can assess and act on risks concerning health and child labour, based on the location of health centres and schools.

In Turkey, data collection under OFIS revealed breaches in labour law. Migrant workers were working illegally long hours during the hazelnut harvest, and this was subsequently identified as a systemic issue in several regions. Equipped with this new insight, Olam responded by educating the farmers on labour regulations and raising awareness of the issue amongst local authorities. Also through OFIS, Olam's hazelnut business has been able to identify potential risks areas of child labour and has set up secure spaces close to the farms to accommodate children of migrant workers during the harvest period to prevent them from working.

Looking ahead, there is huge potential to scale-up the benefits. Olam is exploring for instance, how it can link farmers' phones to a digital wallet and create an entire banking ecosystem through OFIS for people who have been previously overlooked by the financial system. From health and crop insurance to savings facilities and peer-to-peer lending, there is the potential to provide smallholder farmers access to these vital services and meaningfully improve their economic outlook.

OFIS also intends to embrace satellite imagery technology and monitor crop growth, diseases, yields and weather conditions and efficiently liaise back with the farmers for them to take proactive actions. Ultimately, the platform will map all farms and farmers involved in Olam's sustainability programmes and initiatives worldwide, resulting in even more effective targeting of inputs by both Olam and its customer partners, thereby saving costs and increasing both yields and revenue.

At the time it was launched in 2014, there were three game-changing aspects of OFIS when compared to existing technology platforms: the opportunity to incorporate more farmers into the financial system; the level of transparency in the supply chain that could be offered to customers joining the company's sustainability programmes; and finally, the impact that OFIS can have on Olam's own sustainability initiatives. The power of the data collected can now be harnessed in an entirely new way. Although it is clear that crop yields are changing, why they are changing is now known, thanks to a better understanding of the big picture provided by OFIS.

Constraints

The lack of technology and mobile infrastructure in the remote areas where the target group of farmers and their communities are usually based, proved an initial challenge. For this reason, the device was developed with the capability to record data in offline mode to be synchronized later, once an online connection could be established.

Sustainability

OFIS supports Olam's long-term commitment to sustainability. Mapping and surveying each smallholder farm in detail, will – alongside yield improvements – allow Olam to mitigate forest encroachment and other sourcing risks, and is crucial to Olam's ambitious goal for 100 percent of its directly sourced volumes to be traceable and sustainable by 2020.

Replicability

Since 2014, OFIS has been implemented globally across Olam's businesses in cashew, cocoa, coffee, cotton, hazelnut, palm, pepper and rice supply chains, with a footprint of nearly 250 000 farmers and a target to increase this to half a million by 2020.

OFIS is always implemented within the context of wanting to establish a better connection with and understanding of Olam's farmer suppliers. But in order to produce meaningful insight that can deliver impact on the ground, the specific metrics need to be customized for the particular circumstances, risks and challenges of each origin and its producers.

Commenting on the positive impact that Olam's support has had on his livelihood, Muhammed Suleman, a cocoa farmer from Sefwi Medina, Ghana said:

"Before I received my Farm Development Plan I was harvesting seven bags of cocoa … but last year I managed 25 bags thanks to the changes I've made ….The training I have received has showed me I don't need to spray my cocoa so much with pesticides, saving me a lot of money and helping me grow more cocoa … The technology has really helped to bring the world closer to me."[46]

More information on OFIS is available at https://www.olamgroup.com/sustainability/reimagine/olam-farmer-information-system.html

A short video on OFIS is available at https://youtu.be/rOmGouC5Ygc

Contact

Zoe Maddison (Zoe.Maddison@olamnet.com)
Group PR Manager
Olam International

[46] www.bbc.co.uk/news/business-44642175

Case study

3

Mobile solutions, technical assistance and research (mSTAR) project

Introduction

New forms of data and data analytical tools are transforming industrial agriculture and are increasingly being leveraged by a broad set of public, private and nonprofit actors to advance smallholder agricultural production, adaptability, and poverty alleviation in developing agricultural regions. Multiple actors are developing strategies to streamline and standardize data on agricultural trials; data scientists are piloting new methods to estimate yields, growing areas, and even poverty; and research institutions are demonstrating what can be gained by analyzing large amounts of data on crop management, yield, soils, and weather conditions in local farming systems. Exciting platforms such as the CGIAR Platform for Big Data in Agriculture,[47] i2i Data portal,[48] and the Smallholder Finance Product Explorer[49] database, which provides information from 30 service providers in ten countries, and others, as well as projects using predictive analytical methods such as the Aclimate Colombia[50] project showcase the real-life benefits of leveraging data.

USAID funds research and programmes that generate a diverse array of alphanumeric and geospatial datasets related to poverty, markets, agronomic research, natural resource management, nutrition, climate, and much more. Bringing the array and diversity of types of data together stands to unlock significant potential for analysis of relevance to agriculture development programming.

In Nepal, USAID supports 15 Feed the Future (FTF) projects across vegetables, cereals, and lentils value chains. The projects operate in 20 districts across the far-western, mid-western, and western regions of Nepal. In Cambodia, 20 FTF projects operate across fish, horticulture, and rice focus areas. Of these, 14 are part of research initiatives under the Innovation Labs (ILs) operating in Cambodia and six are being implemented by local and global partners. The majority of the projects work in communes and villages in the rural provinces of the Tonle Sap region. All projects currently collect vast amounts of data.

Cambodia and Nepal presented a unique opportunity to leverage an abundance of new and existing data in the service of agriculture development. Yet most of the data are currently stored and managed on local systems and servers. Although data management practices differ across organizations, access to data has been typically restricted to project participants.

47 https://bigdata.cgiar.org/inspire/inspire-challenge-2017/using-ivr-to-connect-farmers-to-market/
48 https://i2ifacility.org/data-portal
49 https://www.themix.org/mixmarket/smallholderfinance
50 http://odimpact.org/case-aclimate-colombia.html

Data are available upon request to other FTF partners; however, no common database or repository existed in the countries where data from all FTF projects are published. With this in mind, FHI 360's mSTAR project supported the Digital Development for Feed the Future (D2FTF) team within USAID's Center for Digital Development, USAID/Cambodia, and USAID/Nepal over the course of 14 months to identify ways to improve the structure, storage, and governance of FTF data and facilitate analysis across the portfolios.

The methodology developed to assess and identify challenges and needs among FTF data producers and users in Cambodia and Nepal is presented below. Building on these findings, we discuss how we evaluated options for adopting a common data structure and utilizing a digital data repository and provide details on steps taken to build consensus on the recommendations and follow-on technical assistance activities to ensure buy-in from partners. Finally, we conclude with several considerations for such undertakings and opportunities for future investments.

Methodology

With an eye on how an open data repository can best support data producers and big data analysis, in collaboration with Development Gateway and Athena Infonomics, we developed a qualitative assessment methodology to identify common priorities, opportunities, and challenges faced by various stakeholders engaged in FTF programming. This was carried out through in-depth key informant interviews with stakeholder organizations in Cambodia and Nepal, and web-survey responses from 30 organizations engaged in FTF-funded programmes and research.

The key findings across the two countries were: there was a reliance on a mix of paper-based and electronic data collection tools; there was a lack of standardized data management protocols and data sharing protocols, including access to catalogues and research findings; and there was some need for training on different types of data analysis. Ultimately, we identified unmet demand among partners and gaps in existing data sharing mechanisms that should be supported by the implementation of a centralized, multistakeholder repository for open agriculture and nutrition data.

Common data structure recommendations

We followed a two-step approach to developing the recommended data structure. In the first step, we studied the underlying structures of the most important datasets in Cambodia's and Nepal's agriculture and nutrition programmes. Sample datasets were gathered from 13 stakeholders – eight in Cambodia and five in Nepal. The datasets were classified as important based on the following criteria: i) their perceived utility to other stakeholders based on inputs from interviews conducted; ii) their relevance to the reporting of USAID indicators; iii) the uniqueness of data contained in the dataset; iv) the frequency of data collection; and v) geographic granularity. This assessment included the following types of datasets relevant to FTF programmes: baseline, midterm, and endline surveys; periodical surveys; monitoring data; and research data from the Innovation Labs. In the second step, we reviewed global good practices for metadata and interoperability standards in open data for agriculture initiatives (such as CGIAR's Open Access and Open Data Support Pack recommended ontologies – AGROVOC, Crop Ontology, Gene Ontology, the International Food Policy Research Institute's Agricultural and Nutrition Technology Ontology, FAO's AIMS portal[51] and the Global Open Data for Agriculture and Nutrition (GODAN) initiative.[52]

Data storage recommendations

In addition to a common data structure, we assessed storage options, which included reviewing six existing digital data repositories along with considering a custom-build repository, and ultimately recommended that the Missions in Cambodia and Nepal adopt an existing storage solution that met their data needs. To ensure the smooth governance and administration of any repository deployed, we recommended that the USAID Nepal Mission and the USAID Cambodia Mission administer the process for setting up and managing the data storage solution and support implementing partners for adopting new standards and processes. Considering the capacity of staff to take on administrative responsibilities, we provided procedures and guidelines on roles and responsibilities, including involvement in data preparation and curation.

Indicator mapping

To supplement our technical evaluation and recommendations, we conducted an indicator mapping exercise in order to show how data sharing, interoperability, and analysis across multiple projects might look in practice. The mapping looked at all indicators and survey questions for FTF projects in Cambodia and Nepal, organized them into thematic areas, identified variables that were common across projects, and highlighted indicators that were similar across projects, but with slight variations. The exercise involved four steps:

- **Collecting and collating all variables from the programme baseline datasets.** If the variable appeared in the project's baseline data, the variable was marked with a "1" in the project's respective column.
- **Organizing variables into thematic areas and subthemes.** This was a subjective exercise that involved organizing the variables based on themes that emerged organically, and further organizing the themes into subthemes. For example, the survey question "How far is your farm from the nearest output market?" was grouped under the theme "Access to Services" and the subtheme "Accessibility".
- **Flagging variables that were similar, but with slight variations**. If two variables are measuring the same thing but are not immediately ready for aggregation because of standardization issues, a "0.5" tag was used instead of "1". For example, the variables "crop calendar for crop X" and "month crop X was planted" were assigned 0.5 tags.
- **Highlighting variables that were common or nearly common across multiple projects.** The total number of similarities for each variable was calculated by tallying the number of 1s and 0.5s tagged across the six projects. This gave a range of "total

[51] http://aims.fao.org/
[52] www.godan.info/

similarities" from 1 to 6, with intervals of 0.5. Variables that had a "total similarities" value greater than 1 were considered common or nearly common across at least two projects.

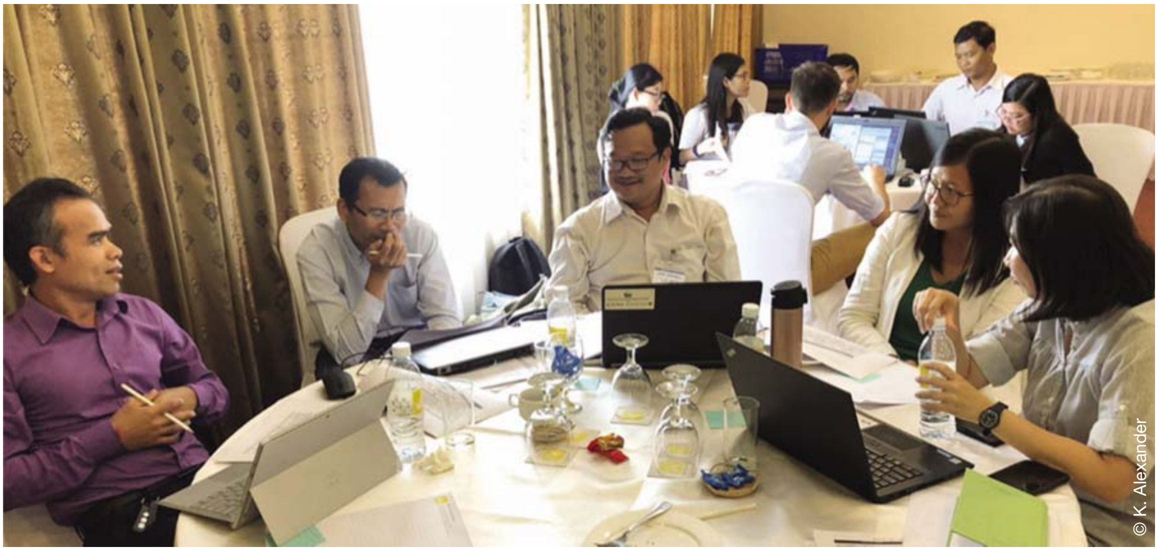
In Cambodia, the six baseline surveys produced over 1 000 variables that were organized into 21 themes and 78 subthemes. In Nepal, the five baseline surveys produced over 1 100 variables that were organized into 16 themes and 88 subthemes. These themes and subthemes were mapped to each other, and variables were tagged to the themes to reveal commonalities across indicators and datasets. Ultimately, we found that few variables were common across the FTF programmes, given the complementary but distinct focus areas of FTF projects. For example, in Nepal we found that the majority of variables had zero commonalities with other projects and were unique to their respective baseline survey. Similar variables tagged to the common thematic areas were typically related to finances and demographics, which is unsurprising given the type of household information commonly collected in baseline surveys.

Impact

To validate these findings and recommendations with Mission staff and FTF partners, and build a consensus on the next steps and proposed technical assistance activities, we carried out "consensus-building workshops" in Nepal and Cambodia. These workshops introduced various data management best practices, demonstrated the use of the recommended open data repository and identified opportunities for collaboration through a common data structure. To facilitate the adoption of new processes that could make the most out of data shared, we also held in-depth discussions of process changes, capacity constraints and other implementation challenges with FTF partners. Overall, participants had a better understanding of many data management concepts after the workshop, performing considerably better on the post-workshop survey: average scores increased by 19.1 percent.

During the validation, it emerged that an institutional data platform, the Development Data Library (DDL), would support in-country data sharing and meet reporting requirements. The follow-on technical assistance focused on: 1) user testing the DDL platform — how to search, upload, browse, signup, create data assets and sets, and view a visualization and geospatial dataset; 2) data management best practices — preparing datasets, metadata, and documentation (codebook, questionnaire, methodology, informed consent, risk analysis, and relevant reports and articles, etc.); and 3) indicator standardization, collaborative research and analysis — by looking at commonalities and opportunities for standardization across partner datasets.

The results of the indicator mapping exercise were also shared with FTF partners during the follow-on workshop. This included looking at baseline datasets across the projects and in both countries to determine: what opportunities exist for cross-portfolio collaboration; what data standardization are needed to facilitate this analysis; and how the DDL, common ontologies, and what best practices in data management can support data sharing. The exercise itself, followed by the hands-on nature of the workshop, was successful in:



Participants at a consensus-building workshop in Cambodia

- identifying variables that partners have in common;
- exploring opportunities for variable standardization;
- understanding which thematic areas have the most overlap;
- demonstrating how data from projects with different outcomes and activities can still provide insights on common outcomes and key research questions; and
- proposing a practical approach to leveraging collective data.

Innovation and success factors

Building on the findings of the stakeholder interviews and data deep dives, we recommended practical guidelines and actionable steps to both partners and USAID missions for adopting a common data structure. A major consideration of the recommendations was selecting ontologies that incorporate country-specific characteristics, including different types of datasets. Resources developed for partners included templates for codebooks, standards for dataset structures uploaded to any open repository, and best practices for data quality – all of which were validated together with FTF partners and USAID missions through the workshop process.

In terms of recommending an appropriate data storage solution, we prioritized identifying tools and governance processes that minimized any extra workload on partners. This also included focusing on what steps missions would need to take in order to ensure buy-in and sustained use from partners: creating ontologies for improving interoperability, investing in data preparation tools, and using their convening power to encourage partners to use data in their programming.

The indicator mapping exercise provided a real-world example of how data management and sharing best practices can facilitate cross-portfolio analysis, using partners' own data to answer the common research questions that they had identified in the first workshop. In addition to the insights drawn from this activity, the Excel database itself was as an important output of this exercise. At the follow-on technical assistance workshops, partners and mission staff were given the database and encouraged to explore it for overlapping themes – for example, by using filters and pivot tables to draw their own insights related to opportunities for standardization (Figure 18).

	A	C	E	AC	AD	AE	AJ	AK	AL	AM
1	Theme	Sub-Theme	Full Indicator	CSISA Tot	ILRI-Heif	KISAN	NSAF Tot	PAHAL	Total Similarities	W/Similarity?
2	Demographic	Individual	Respondent Name	0	1	1	1		3	1
3			Respondent Sex	0	1	1	1		3	1
4			Respondent Age Years	0	1		0		1	0

Figure 18: Screenshot of compiled database

Constraints

The primary constraints faced were related to designing a methodology for indicator mapping that would be relevant for FTF programmes and useful for designing additional collaboration activities in the future. Ultimately the indicator mapping findings were presented with a few caveats as presented below.

First, only baseline surveys in the FTF datasets were collated. To build upon the usability of this exercise, expanding the scope to include other datasets should be considered. This will allow for stronger insights about where areas of redundancy and potential collaboration exist. In addition, not having access to variable codebooks meant there were limitations in understanding the units, wording, coding, and formatting from project to project.

Second, thematic tags emerged organically, and the organization of variables was a subjective analysis based upon the agreement of two researchers not closely associated with the design or analysis of the baseline surveys. In order to capture the multidimensionality of many variables, a thematic cross table was designed to allow each variable to be tagged under more than one theme (for example, "number of livestock owned" could be tagged to both "livestock" and "assets", instead of just "livestock"). However, because of the complexity of the cross table and its limited value in terms of resolving the issue of subjectivity, the cross table was not included in the final product. In the future, machine learning analysis and detailed coding manuals could allow for a more scientific methodology for thematic assignment.

Third, the original methodology for this exercise included tagging each variable to objectives and outcomes specified in the FTF Results Framework, but because the current results framework does not include the level of detail required to tag each variable or survey question to one (or more) outcome results, the exercise had to rely on subjective interpretation. Additionally, with only three FTF objectives and nine outcome results, such an analysis would have either oversimplified the multidimensionality of variables (by tagging only one outcome to each variable) or overcomplicated the interpretability of the results (by tagging more than one outcome to each variable). For these reasons, the mapping of baseline variables to FTF outcome areas was not included in the final product. Improved coding manuals and results frameworks, as well as agreement between FTF and its partners on the relevance of survey questions/variables to FTF outcome results, could help facilitate this kind of mapping exercise in the future.

Lessons learnt

Big data has great potential to enhance data use and sharing for the purposes of programmatic improvements. However, we found common challenges and needs across stakeholders:

1. Data interoperability – with the lack of formal data-sharing mechanisms, it is important to focus on three distinct aspects: exchange of data between systems; data files that can be interpreted; and data that is understood by the system. This includes:
    a. meeting basic structural hygiene standards that allow for easy integration with other applications and repositories by keeping data belonging to the same dataset within a single file (unless there is a strong reason not to do this); using clear column and row names, proper variable names, and standard coordinate systems for georeferenced files; and using standard codes for representing different types of missing data; and
    b. capturing important metadata elements at the project and dataset levels; and creating corresponding codebooks in a tabular structure, saved in a non-proprietary file format such as a comma-separated values (CSV) file or a tab-separated values (TSV) file.

2. Deploying data standards – for structuring and defining data, it is important to start with basic data management protocols to define the types and properties of variables in a dataset. Collect, store, and report all data using standardized vocabulary and units of measurement, wherever possible and specified in the codebook. For example, by using:
    a. a controlled vocabulary such as AGROVOC[53] – which covers food, nutrition, agriculture, fisheries, forestry, and environment – and others such as the Global Agricultural Concept Scheme (GACS), the National Agricultural Library Thesaurus (NALT) and the Centre for Agricultural Bioscience (CAB) Thesaurus offer sector specific vocabulary lists; and
    b. standard units of measurement and naming schemes for country-specific variables that adheres to context and local governance, for example using appropriate naming conventions for agriculture variables and administrative units and agreed-upon units for reporting area, length, weight, volume, and dates.

[53] http://aims.fao.org/vest-registry/vocabularies/agrovoc

3. Data management and storage – As research and data management practices have become more transparent, digital data repositories have also advanced beyond basic functions to support data upload and storage. Assessments should draw on the resources and good practices that have emerged for uploading, storing, and sharing data, including guidelines created by the Data Curation Centre[54] and Data Seal of Approval,[55] and the FAIR principles (making data Findable, Accessible, Interoperable and Reusable).[56] We recommend that any data repository include the following important functionalities:
    a. the ability to assign unique identifiers, restrict access to certain data, apply reuse licenses, be discoverable by search engines, provide interoperability through application programming interfaces (APIs), and allow any file type to be uploaded and shared; and
    b. rich and granular metadata fields, advanced search functions, customized use metrics, auto-generated "preservation" file format, built-on integration for data analysis and visualization, and links to other data catalogues.

Contact

Nina Getachew ngetachew@fhi360.org
Kathryn Alexander kalexander@developmentgateway.org

[54] http://www.dcc.ac.uk/resources/how-guides-checklists/where-keep-research-data/where-keep-research-data
[55] http://www.dcc.ac.uk/resources/curation-reference-manual/chapters-production/audit-and-certification
[56] www.nature.com/articles/sdata201618

Case study

4

Leveraging satellite data and artificial intelligence to drive financial inclusion for smallholder farmers

Introduction

Globally, there are 500 million smallholder farmers, yet less than 10 percent have access to credit to buy improved inputs or technologies that could improve their yields, incomes and food security. Without financing, farmers lack the means to improve the quality and quantity of their production and remain trapped in a cycle of low production and poverty.

In Myanmar, there are nearly eight million smallholder farmers but less than 5 percent of the population is considered "bankable." Banks do not lend to smallholder farmers because they exist largely outside the formal economy and are perceived as too risky since they lack collateral – only 32 percent of Myanmar's farmers have property deeds or bank accounts and transaction histories that are typically used to assess borrowers. Their remote locations increase the challenge and costs of collecting relevant data.

More recently, companies have begun to leverage alternative data such as social media activity and digital footprints to generate credit scores for unbanked populations. But these platforms target urban populations with short-term, consumer loans. They effectively exclude smallholder communities since farmers do not have significant digital activity and need larger, longer-term loans for productive purposes.

Methodology

Harvesting Inc. (more generally known simply as Harvesting)[57] leverages satellite data and artificial intelligence (AI) to provide timely, cost-effective, and accurate insights into the crop activity of individual small farms (less than a hectare) worldwide. Currently, Harvesting delivers crop data to clients through agri-lending software that is designed to help reduce the information asymmetries and transaction costs of serving smallholder farmers. A mobile data collection app allows on-site data collection, including the farm's geo-coordinates, at the farm. Digitization improves the speed and accuracy of data gathering compared to manual methods. Collected data flows into a credit risk scoring system that incorporates alternative data, including Harvesting's remote sensing data on crop activity, to generate credit scores for borrowers. Credit scoring can also improve the efficiency of the loan approval process. Once a loan is made, farmland monitoring enables lenders to monitor a farm remotely and cost-effectively and intervene with borrowers pro-actively.

[57] www.harvesting.co

By reducing the information asymmetries and transaction costs of working with smallholder farmers, Harvesting seeks to enable farmers to have access to financial services.

The goal of the practice is to increase the number of smallholder farmers who have access to loans and to increase the loan amount for farmers who are "good borrowers". A pilot project began in October 2018 and is still in progress. Once loans are made through Harvesting's credit risk scoring model and monitored through farmland monitoring, the company intends to track metrics including: number of smallholder farmers (SHF) receiving loans; amount of loans (in USD) made to SHF; and percentage increase in loan amount to farmers who are "good borrowers". The scope of the pilot project does not include assessment of changes in the livelihoods of SHF.

As more data about more farms in the region become available, Harvesting would like to offer more predictive analytics and decision-making tools for individual SHF in the long-term, such as recommended planting and harvesting dates based on the data currently available on each farm and prevailing weather conditions.

The pilot project is still in the implementation phase – Harvesting has not yet begun to track metrics – so it is too early to judge its success. Early indications are positive though: (i) the credit risk model was back-tested by Harvesting and the company was able to match historical repayments with high accuracy; and (ii) the automated credit risk scoring is faster than the client's existing approach which is not automated.

Harvesting's solution is unique because no other provider can, first, generate data on the crop activity of individual smallholder farms cost-effectively and accurately, and second, deliver it to financial institutions through software designed to help decrease the transaction costs of working with farmers. Although there are other organizations using satellite data to provide analytics on agriculture, they are not focused on agriculture and work primarily in developed countries. Harvesting focuses exclusively on smallholder farmers in frontier markets. Moreover, Harvesting's software is designed to enable lending to smallholder farmers. Other companies use alternative data such as social media and digital footprints to derive credit scores to reach under-banked communities, but they primarily target urban populations with consumer loans and effectively exclude farmers who have little or no digital footprint and who need longer-term, productive loans.

Lessons learnt

Given the value of data in the services offered by the company, the quality of data input is critical:

(i) Harvesting learned in this pilot experience that the company needs to prepare its clients better, and provide more specific instructions on the types of information needed and how to collect it;

(ii) even with better preparation, there will likely be translation errors and other unexpected issues, so it is important to build in time for making sure both parties understand what each specific line item means and scrubbing the data; and

(iii) obstacles and issues may come up in ground data collection such as during the rainy season or because of the lack of motorized vehicles and may slow down the time it takes to collect data, so there is a need to prepare for such occurrences in advance.

Most of the challenges encountered in the pilot project were a result of data and IT problems.

Sustainability

Harvesting's current pilot project is funded in part by FMO, a Dutch development bank, to facilitate the adoption of, and de-risk the investment in, a new technology by the project partner, Maha Agriculture Microfinance. Once the project is implemented, the expectation is that Maha will continue to pay for the service on a commercial basis – thus the practice will be sustainable.

Because Harvesting's solution is designed to benefit SHF, it is also designed to be economically viable. There is no reliance on high-cost private satellites, only publicly available data from the satellites of the United States' National Aeronautics and Space Administration (NASA) and the European Space Agency (ESA), as well as other public data sources, are used. Also, because Harvesting's solution can both expand the client base of service providers to include SHF, and reduce the transaction costs of working with them, the company's software generates a positive return on investment such that clients would likely be willing to pay for it on a commercial basis.

Replicability

Harvesting is currently building crop models and testing its products in select markets in Africa and Asia and would like to do more pilot projects in these regions. India is seen as a critical testing ground given its diversity of agro-climates and crops as well as its sheer number of smallholder farmers. The more farm data points that are gathered, the more the company's AI models learn and can make recommendations about farm-level needs.

Harvesting's satellite imagery processing pipeline has been designed and built to be scalable and able to process petabytes of data efficiently. However, some "adaptation" is needed: every time new data sources are added new application programming interfaces need to be developed and each time a new agro-climatic region is entered, crop models need to be calibrated and validated for that new region.

Contact

Julie Cheng
julie@harvesting.co

Case study

5

Delivering remote flood analytics as a scalable service

Operationalizing satellite-based and near real-time flood monitoring for local emergency response in the Republic of the Congo

Introduction

A community's ability to absorb a shock and prevent a flood from becoming a disaster is key to its long term resilience. However, governments, communities and other actors can only reliably reduce the number of deaths and protect their economies if they know where vulnerable people and assets are when a disaster occurs or looks likely to occur. For flood preparedness, the identification of people and assets most exposed to flooding enables government and aid agencies to prepare aid and plan a response, as well as rezone assets and design protective infrastructure.

Each year humanitarian relief meets only seven percent of economic losses from disasters in the developing world. With floods doubling worldwide (because of climate change and population migration), increasing aid after a disaster is only part of the solution. It is critically important to significantly decrease the estimated USD 29.9 billion losses that occur as a result of flooding every year by enabling vulnerable countries to reduce their risk and address impacts more effectively. This starts with information and the ability to use it. If governments and other responders do not have locally relevant information about a flood, they cannot prepare for it or respond effectively when it happens.

Traditional flood mapping is manual, slow and expensive – it requires locally gathered and calibrated models for high accuracy, sometimes taking years to map a stretch of river, and can cost millions per project. However, C2S's flood analytics and monitoring from satellites, produces high resolution and near real-time monitoring and is fast and accessible even to users with limited risk modeling training and unreliable Internet services.

C2S's platform harnesses dozens of global satellites and community intelligence to detect disasters as they happen, predict elevated flood risk situations, and provide flood vulnerability maps at a fraction of the cost and time of traditional methods. Customized versions of the platform that integrate the science of local flood dynamics are designed to support flood response and long-term preparation for a country's government. First, near real-time flood monitoring alerts the government to flooding anywhere in the country and rapidly pinpoints people affected during a crisis and helps in delivering the disaster relief needed. With distilled short offline messages and interactive decision support, managers can quickly share damage reports with everyone from international agencies to first responders. Second, calibrated flood triggers, based on local flood and weather dynamics, warn users of potential flooding before it happens. Third, flood frequency and probability maps based on tens of thousands of satellite images over 35 years reveal areas of risk and can be updated with a click of a button. These analytics enable authorities to keep people and assets out of flood prone areas, design dykes and dams, identify wetlands, and more.

C2S has provided flood analytics to over nine countries, with a plan to be in 50 with over 200 projects by the end of 2022 through partnerships with existing insurance partners, governments and aid agencies.

The following report is an example of a recent pilot project in the Republic of the Congo between the government, the UN World Food Programme (WFP) and Cloud to Street. UN WFP worked with C2S through its Innovation Accelerator programme which supports and scales high potential solutions to hunger worldwide. The WFP Innovation Accelerator programme identified and utilized C2S's service for their humanitarian response activities by leveraging C2S's transformative technology.

Context

In November 2017, the city of Impfondo in the Republic of the Congo experienced a serious flood event, leaving 5 000 people in need of food assistance. However, the WFP did not learn of the flood for an entire month after it occurred, and once they did receive information about it, information about the size of the flood and the need for food was unclear, thus delaying the response. Alerts about the flood initially came by word of mouth, and later from field staff deployed from the capital.

The goal of this pilot project was to demonstrate the value of flood information services based on satellite imagery for monitoring floods. C2S sought to assess whether or not satellite imagery could deliver useable and impactful data on flooding to WFP and to the government ministries of the Republic of Congo. For WFP, this would mean reducing response times by rapidly assessing if, where, and how much food relief was required. For the Congolese government, this would mean providing information on flooding located in remote parts of the country where there is little regular contact. The goal of both stakeholders was to improve on the benchmark set by the November 2017 Impfondo flood event.

C2S proposed testing the impact of its near real-time satellite flood monitoring service on three critical disaster emergency activities in the country: 1) food relief by WFP; 2) government response to floods; and 3) coordination between stakeholders during a crisis.

C2S designed and implemented a locally customized automated flood and rainfall monitoring online tool, provided as a service within an interactive dashboard for government and WFP users. Government users, based in Brazzaville, included the Ministry of Social Affairs and Humanitarian Action, the Meteorology Office, and the HydroMet Office, among others. The system scientifically optimized and combined global flood detection algorithms to the unique flood dynamics of the region to generate flood analysis and reports that can be verified from the ground and shared through WhatsApp alerts daily.

This service has three main features:

1. a user-centered dashboard and offline tools customized for the most important local flood needs based on what information can reduce flood vulnerability;
2. flood information leveraging the best available science, satellites, and community intelligence; and
3. day-to-day support and local capacity building to make sure users understand the data and can use the information to make decisions.

Cloud to Street's approach to flood detection and response is shown in Figure 19.

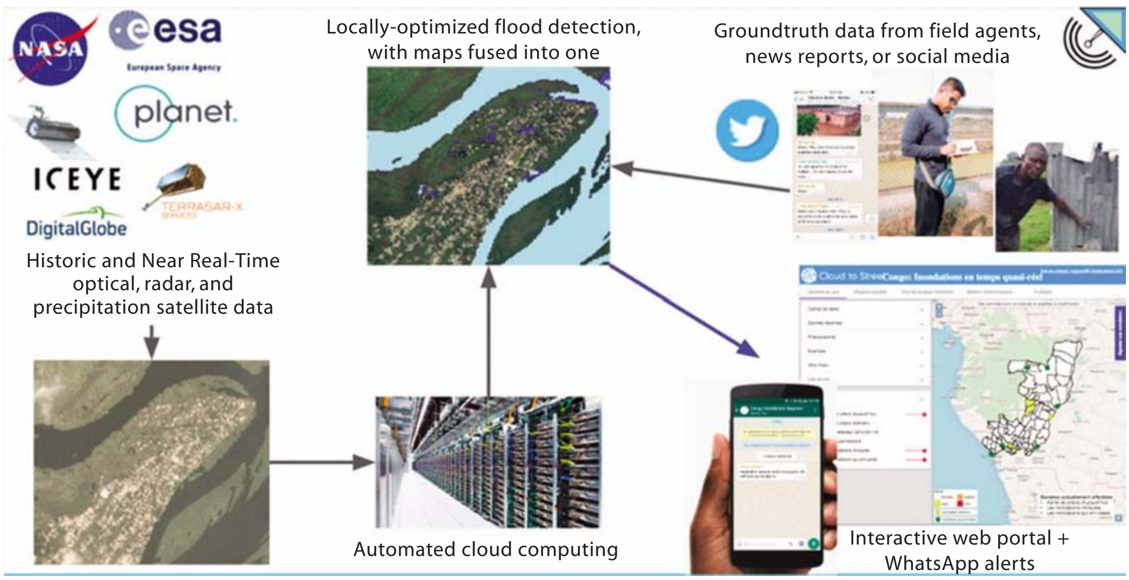


Figure 19: Cloud to Street's approach to flood detection and response

Starting with optical, radar, and precipitation satellite data, and leveraging automated cloud computing and groundtruth data from field agents, news reports, or social media, C2S perform locally-optimized flood detection. The results are then presented in an interactive dashboard combining flood analysis and reports (Figure 20).

"A" shows the "recent data" page, which contains a map of precipitation alerts by district, with locations of observed flooding events throughout the country. Collapsible groups of layers on the left allow for the display of different precipitation, flood, and contextual information. The "Report a flood" button in blue (not shown here) allows users to manually report a flood to C2S. "B" shows the "current situation" page, which summarizes the latest information on each flooding event using satellite imagery, flood model results, and rainfall levels. "C" shows the "meteorological bulletin" page, which provides monthly and ten-day summaries of rainfall throughout the country compared to historical averages. "D" shows the "daily rainfall from the HydroMet Office" page providing automatic daily precipitation data for two weeks before and after the current date, allowing power users to download tabular data for use in their own reports and analyses.

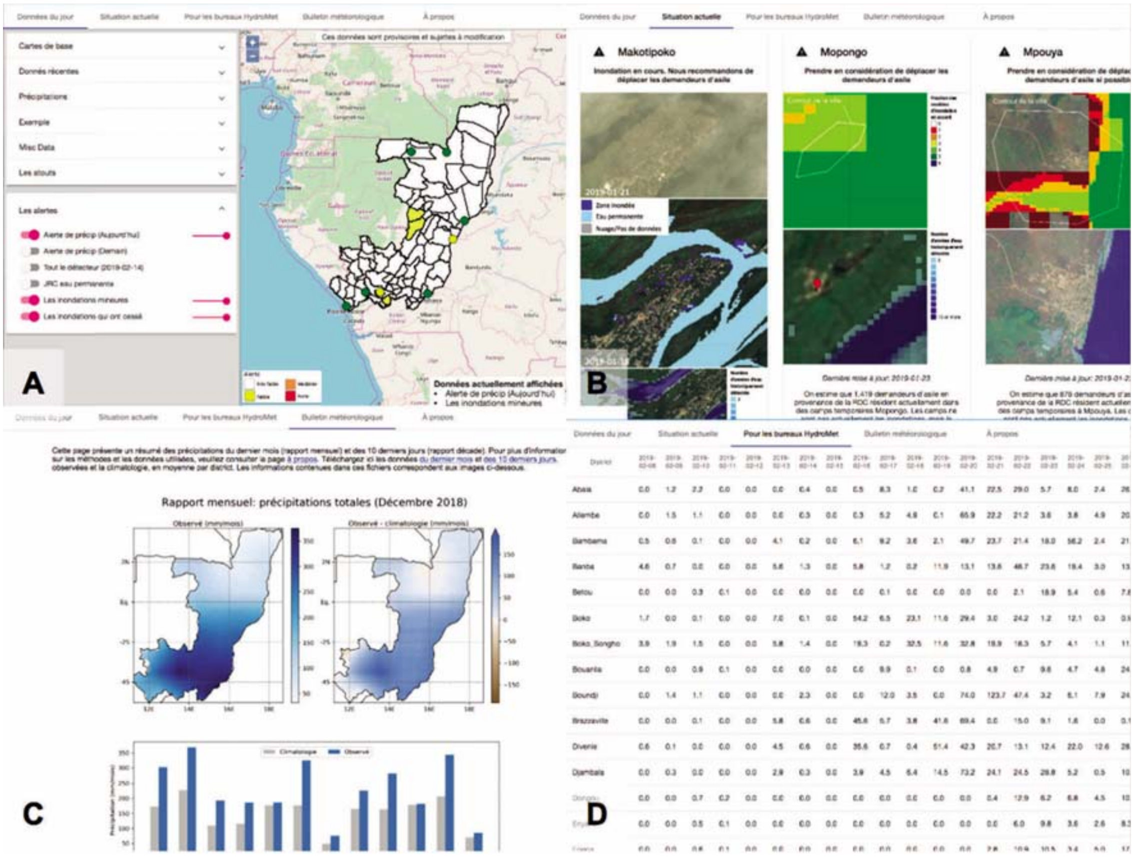


Figure 20: The main pages of Cloud to Street's near real-time flood monitoring dashboard

In operating this system, C2S created rapid flood maps designed to be useful for decision making, initially with public satellite imagery and then also using commercial satellite imagery when public satellite imagery was not sufficient for providing useful information on flood occurrence.

Overview of findings

Over the course of monitoring, the C2S system identified eight flood events and assessed the flood risk of four additional sites with asylum seekers. Five of the eight events were identified using public satellite tools and three were reported by local stakeholders in urban areas. These floods impacted 33 homes in Makotipoko, 26 homes in Mossaka, at least 11 homes in Nkayi, and also identified flooded roads and the risk of larger potential flooding in Ouesso and Sembe.

The flooding event in Mossaka was the strongest example of the value of timely flood information. Flooding was first detected by Sentinel-2 satellite imagery in the town of Mossaka on 9 November, and parts of the town continued to flood intermittently until 18 January. In addition to the Mossaka flooding event, C2S's system detected minor or potential flooding in three other towns (Nkayi, Ouesso and Sembe). C2S communicated these situations to a WhatsApp group and recommended contacting local field agents for confirmation.

User design exercise for local stakeholders

Cloud to Street used human-centered design methods, customized for flood decision making in the developing world, in order to assess the true needs and capacities of its users. Starting with these user design methods – not prescribed technological solutions – helped ensure that we provided what the end users wanted and that they could unlock the value of what they got, and that the solutions would therefore inspire long-term use of the tool. Exercises centred around stakeholder mapping and timeline analyses for recovery efforts and extensive scenario discussions to ensure a usable design product.

Overall, the process proved essential for identifying key local stakeholder groups, clarifying the chain of command, identifying where information could change the flood process in places, and profiling the current capacity in general.

Outcomes

1. Flood monitoring for asylum seeker sites

On 28 December, Cloud to Street was made aware that about 17 000 asylum seekers from the neighbouring Democratic Republic of the Congo had crossed the border and sought refuge in several sites along the Congo River on the Republic of the Congo side. The UN Refugee Agency (UNHCR) was concerned about the flood risk of these sites, and sought information from external sources. C2S mobilized quickly, providing an initial briefing on the flood risk based on historical flood patterns by 29 December.

By 3 January, additional information from six flood models was provided for a more comprehensive assessment of flood risk. At that point, daily checks of the four main asylum seeker sites to the monitoring process were included in case flooding occurred; this also included providing daily information on “current situation” cards for local decision makers. On 16 January, these recommendations were presented to UNHCR formally, which was then conveyed by them to local government actors on the ground.

On 8 February, UNHCR reported that the government agreed to relocate refugees from the highest risk site (Makotipoko) to one of the sites with lower risk (Bouemba). Makotipoko was also the site with the largest number of refugees and this timely evacuation averted an impending disaster. For C2S, this represented the successful deployment of critical satellite information in a way that spurs quick interventions from the government and aid agencies.

2. Building coordination and capacity through WhatsApp

The system was designed to provide useable flood information for government users to incorporate into their decision-making process. The initial assumption had been that government users would check the pages of the dashboard regularly, or would check the dashboard when alerted. However, as the pilot project proceeded, it was found that more direct communication through WhatsApp allowed for greater uptake of flood alert information based on monitoring web visits to the dashboard from Brazzaville using Google Analytics. Therefore, a local WhatsApp group with representatives from different government ministries and other local stakeholders was created and subsequently daily summaries of the "current situation" page were sent to that group. This allowed for confirmation from users on the ground about events observed with C2S's system. The WhatsApp group served as a focal point of coordination among ministries who had previously lacked a common source of information about flooding.

3. Building technical capacity through training

Cloud to Street added to local capacity of the government ministries through training at the pilot project's onset in October, through monthly updates starting in December, and through shared coordination around flooding using the WhatsApp group. On 2 November, 18 officials from seven government ministries attended a training event on the C2S System. During subsequent monthly conference calls attended by stakeholders, they received new information from the system. In total, 28 local stakeholders were trained on how to use the service, with 18 from local Congolese government offices and ten from non-governmental agencies such as WFP.

Challenges to flood resilience in the Congo

1. Coordination between stakeholders

Lack of clear coordination between various stakeholder groups made it difficult to assign responsibility clearly to specific ministries during a serious flood event. It is noteworthy that during a fact-finding session (part of which involved groups of participants in creating maps) initiated by C2S prior to service implementation (Figure 21), each group drew a stakeholder map to communicate their understanding of which ministry was responsible during a flood emergency and there was no consensus on the chain of command or process.

"A" shows a prototypical stakeholder map drawn by a government ministry user communicating the chain of command for responding to floods. The blue circle shows the Ministry of Social Affairs at the centre of other actors, something disputed by that particular ministry.

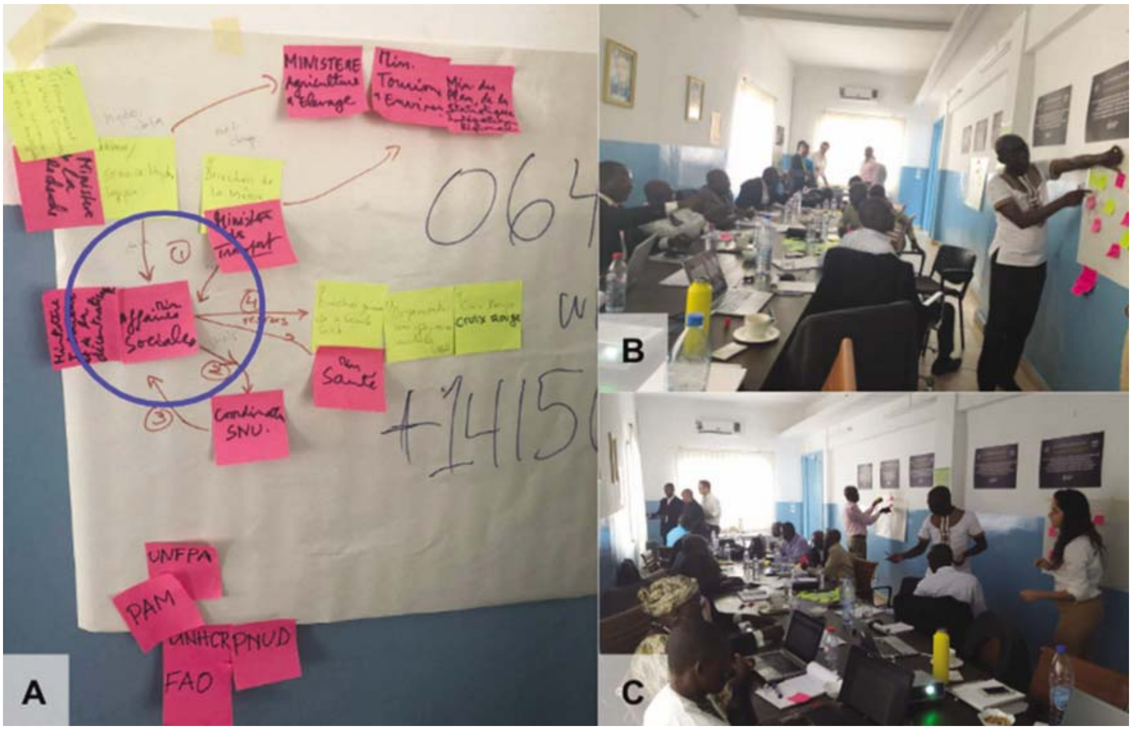


Figure 21: Communication capacity for flood emergencies in Republic of the Congo

2. Data availability

Compounding the coordination problems among users was the limited tools that could be utilized on flood issues. As the Republic of the Congo is a country that has experienced conflict over the past several decades, only approximately 13 out of the more than 80 historical hydrological gauges are currently operational. Rainfall data is also limited, with information recorded manually and reported once a month from over 15 points around the country (Figure 22).

Notes: (A) Government official from the HydroMet Office comparing the rainfall point data used locally to satellite rainfall data. (B) The reduced number of gauges because of past conflict (top, historical gauge locations; bottom current gauge locations). (C) The manual data collection process for collecting rainfall data.

3. Technical capacity for using satellite-derived information

Beyond the limited tools available, there was also limited capacity for understanding and using publicly available tools that could improve existing processes. Few government ministry officers have staff with appropriate science and engineering training to understand how to use data from satellites, with only the HydroMet Office having this technical capacity. Moreover, slow Internet speeds and low bandwidth further limit the ability to analyse and process such data.

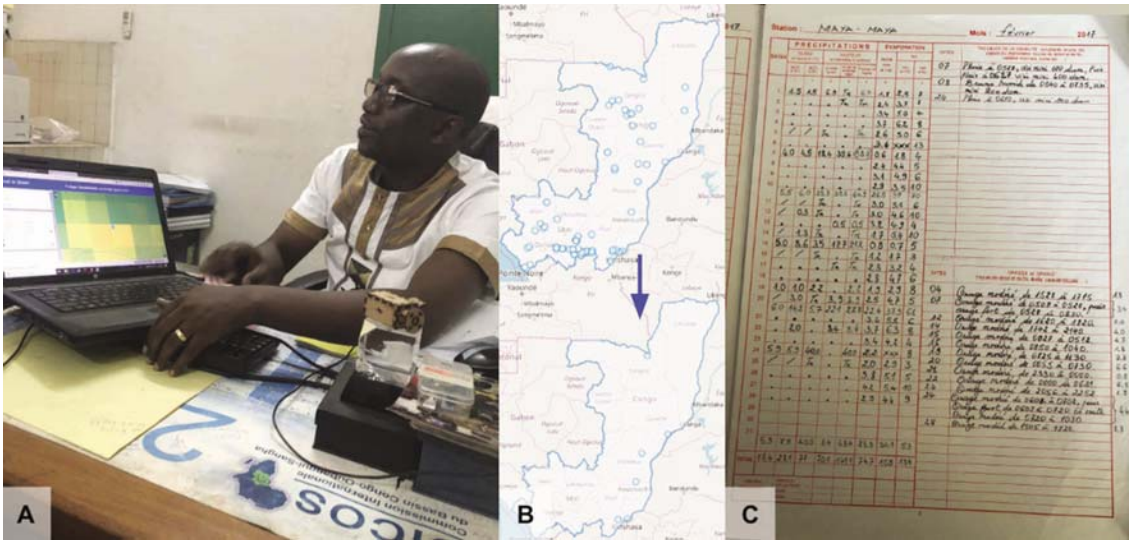


Figure 22: Existing flood information limitations for Republic of the Congo

Conclusions

1. C2S's flood monitoring services identified floods and impacts earlier than would have been identified otherwise by traditional services.
2. C2S was able to respond rapidly and within days to an urgent situation around flood risk for 16 000 refugees.
3. C2S created and facilitated a local WhatsApp group with representation from about a dozen relevant government ministers who used the group to share information and receive daily updates during flooding events.

Contact

Email: support@cloudtostreet.info

AtSource – Connecting customers to the source of supply

Introduction

Food systems are coming under increasing pressure from the growing population while at the same time the earth's capacity to continue to provide water, nutrients and a stable climate is at a tipping point. Moreover, understanding the implications of agricultural products for sustainability is increasingly of concern to consumers worldwide. Millions of farmers are also in need of a future that is economically and socially sustainable, one in which future generations will invest and grow.

However, reliable and consistent data and insights required to address these issues are difficult to obtain because of the high level of fragmentation in global agricultural supply chains, particularly in emerging markets. Crops often make their way from farmers in rural villages, change hands through many intermediaries, and are combined with other volumes before reaching a manufacturer such as a food company. This makes traceability difficult. What is required is a new set of tools to inform and plan social and environmental initiatives at the scale and urgency required to benefit farmers, rural communities, consumers and the planet.

The development of AtSource stems from Olam's purpose, to "Re-imagine Global Agriculture and Food Systems" and builds on the many existing initiatives to secure sustainable practices across Olam's operations and its third party sourcing. Such initiatives include the Olam Livelihood Charter, which today embraces approximately 445 900 smallholders across Africa, Asia and South America. In Olam's third party supply chains, the company has been rolling out its Supplier Code – an agreement that all Olam's suppliers of raw materials and products sign as a commitment to abide by Olam's key sustainability requirements on forest conservation, child labour, human rights, amongst others, as well as international guidelines.

Since 2014, Olam has been gathering data at farmgate level through the Olam Farmer Information System (OFIS), to provide personalized advice to farmers on improving farming practices, on traceability and on financial access. AtSource builds on this digital mapping capability by providing information on traceability for each and every stage of a product's journey – from the farm, through logistics and processing, up to the customer's door.

AtSource is the culmination of Olam's extensive social and environmental expertise, channeled into a transparent and customized digital sustainability solution.

Methodology

The rich data provided on the platform is grouped into three pillars – Prosperous Farmers & Farming Systems, Thriving Communities, and Regenerating the Living World, and is presented on a customized digital dashboard. It offers three tiers with progressively enhanced levels of impact, enabling customers to choose their level based on their priorities and sustainability journeys.

The first-tier level is AtSource Entry. This provides customers with reassurance that suppliers are engaged in responsible sourcing principles and practices under the Olam Supplier Code. They also have access to a unique Eco-calculator, which enables them to access, in an instant, the environmental footprint of their sourcing at country-product level, alongside risk ratings on key environmental and social topics provided by credible third party organizations.

At the second level, AtSource Plus, Olam works with its customers on various sustainability metrics, which measure areas such as gender equity, education, market access, fair pricing, greenhouse gas emissions and water use. Customers access this information on a customized digital dashboard so they can see the impact of the total journey of their product from source. They can also look across and compare the different products they source from Olam. With these insights, they are able to make more targeted decisions and take efficient action to help make their supply chains more resilient to issues such as climate change.

The third and top tier, AtSource Infinity, is where Olam wants to use the data to co-create sustainability programmes with its customers – in other words, Olam wants to deliver a net-positive impact on communities and the environment and do so at scale.

Impact

In line with the ambition of AtSource Plus, Olam's sustainability programme in Viet Nam engages 1 245 coffee farmers to ensure positive social, environmental and economic impacts. The programme will provide Olam's international customers with 100 percent traceability to the farmer groups that produce their coffee products in Viet Nam. Transparent data and farmer insights offer opportunities for more precise and resource-efficient sustainability programmes, and supports their commitment to source all their coffee from Rainforest Alliance Certified™ farms.

As a result of AtSource, Olam's coffee businesses have institutionalized a higher level of third-party supplier engagement across all origins with the Olam Supplier Code and have introduced measures to ensure internal verification of compliance.

On request, Olam's spices business in the United States of America is now able to provide a customer with the total carbon dioxide emissions of their finished onion and garlic products, from cultivation, through to processing, to support any company's report on their Scope 3[58] emissions assessment as required by the Greenhouse Gas Protocol.

[58] Scope 3 concerns indirect emissions.

In the short term, the benefit of AtSource will be in bringing a degree of coherence and consistency to how Olam engages with producers and understands the challenges they face. As the data legacy grows, it is anticipated that the insights provided by AtSource will attract investment finance for project co-creation and that projects will increasingly be tailored in response to learning about what interventions have been most successful.

There were a number of factors that led Olam to introduce AtSource. Olam has a heritage of leadership in programme development and implementation, from the Olam Livelihood Charter, to multistakeholder jurisdictional programmes, all driven by its purpose, to “Re-Imagine Global Agriculture and Food Systems”. Ultimately, Olam’s business depends on sustainability to continue to secure volumes for its customers.

The AtSource platform leverages emerging market interest in product origination and has created a way to show how farmers, customers, and consumers can derive value from data and insight. It also provides Olam with a platform to showcase its sustainability efforts and differentiate the company as a supplier that can offer 100 percent traceability on its products. This new data and insight evidences Olam’s claims of sustainability leadership and incentivizes long-term strategic partnerships based on transparency and trust.

Additionally, there was an opportunity to capitalize on Olam’s recent investments in digitization – from on-the-ground data capture and analysis to building a cross-business digital infrastructure. It is often said that data is the new currency, and Olam aims to be at the forefront of developments to bring data, insight and transformation to global supply chains.

The integrity of the AtSource offering is supported by core metrics that allow a customer purchasing multiple products to have a common dashboard. It also caters to the customers’ desires for tailored insights and solutions by allowing for additional product and origin specific metrics. From a broader perspective, AtSource has been designed to have relevance to global initiatives, such as sustainability certifications and reporting frameworks.

Constraints

In terms of the development of the AtSource platform itself, the complexity lay in developing a set of core metrics that allow for differences in large and small supply chains and apply cross-commodity. For example, in developing countries, smallholders may only farm a few hectares, whereas the majority of onion and garlic farmers in Olam’s spices business in the United States of America are well-diversified, growing many crops with an average size farm of 800 ha or more. Other differences to consider include the high level of mechanization, use of precision agriculture and the stringent regulatory environment.

There were also challenges in operationalizing AtSource and getting all parts of the business to work towards a common goal. From agricultural operations, to manufacturing, engineering, supply chain management, to information technology (IT) and human resources, the offering needed to be well understood at all levels. Data and other IT systems needed to be mapped and merged, whereas processes and methodologies for data collection and validation needed to be standardized.

The ongoing challenge however, is gathering the core metrics and footprinting input data from the origins and validating it to ensure its accuracy, plus creating the digital infrastructure to publish robust data on the platform.

Data gathering and validation in smallholder landscapes is complex and resource intensive and there are substantive challenges in reaching beyond the farm-gate. The numbers are daunting: there are 4.8 million farmers producing Olam products and each of them form part of wider families and communities affected by agriculture. However, AtSource provides an opportunity to get ahead of a critical trend and create a market advantage for the farmers and for the company.

Farmers are not charged for providing data for AtSource. In fact, Olam has invested substantially in the human and technical capital needed to bring AtSource to market. The company believes that farmers and the natural environment will benefit materially from the insights provided by AtSource, as these will inform better investments and more tailored actions than have been hitherto possible. Continual improvement is at the heart of AtSource and over time the company will learn what works well and where efforts should be prioritized.

Building on what was learned from the test and learn phase, Olam has now broadened the AtSource offer to include 20 new coffee origins, onions from Egypt, cocoa from Ghana and cashews from Ghana and Mozambique. This list will grow over time as origins build the data and sustainability capacity to enter AtSource.

Contact

Zoe Maddison
Zoe.Maddison@olamnet.com

Case study

7

WAGRI – the agricultural big data platform

Introduction

Recently, agriculture in Japan has faced some serious difficulties, mainly because of the rapidly declining number of farmers. This has meant the loss of tacit knowledge of farming passed on from generation to generation and the failure to cultivate deserted but cultivatable lands. In fact, the area of deserted cultivatable lands is well over one percent of the land area of Japan. Faced with this situation, a data science approach is expected to bring about effective alternatives for both tacit knowledge and manpower. In fact, a data science approach to problem solving is not really a novel one as a well-known example can be found at the end of nineteenth century, when William S. Gosset, an employee of Guinness & Co the Irish brewer famous for its draught beer, developed a t-test to estimate the taste of a large volume of beer based on small samples of ingredients. Since then, various data science approaches have been tried in the area of agriculture. These trials have not had much impact on farming in Japan, particularly as the cost of data management has been rather expensive. The loss of tacit knowledge and the experiences of individuals has weighed heavily in Japan and the situation has become more serious in recent years. However, the rapid cost reductions of computing resources has brought about unprecedented opportunities to realize data-driven agriculture.

Based on such circumstances, the possibility of constructing a big data platform for agriculture has caught the attention of farming organizations, private enterprises such as farm machinery manufacturers, local government bodies, the Ministry of Agriculture, and the Cabinet Office. WAGRI, a pioneering big data platform for agriculture, was developed initially in 2017 in response to the Prime Minister's statements at the Cabinet conference on Japan's Future Investment Strategy.

The next section describes the general characteristics of agricultural data and the issues for constructing a big data platform accompanied by the introduction of data implementation on WAGRI. The section that follows this describes the architecture of services for users including the user-defined application programming interface (API) and the architecture of access control. Then there is a section that introduces the possibility of discovering new agricultural knowledge based on the big data, and this is followed by the conclusions.

Characteristics of agricultural data and implementation of WAGRI

Among the most important items of agricultural data are those related to crop health such as climate, soil ingredients, and pests. Most agricultural data are locally collected by different organizations, so they are widely scattered across numerous organizations and locations. Furthermore, different definitions of data items are used along with different file formats such as CSV and txt. Additionally, they are updated randomly and frequently redefined without any advance notice. As such, the data are so diverse that it is impossible to develop databases strictly defining the individual attributes of each column of data such as number of digits, alphanumeric property, and relations to other columns. These factors make the creation and application of an agricultural big data platform a great challenge. To cope with this, WAGRI has adopted a flexible data structure to avoid strict definition of each item of data.

JavaScript Object Notation (Json) has been widely adopted as the data format for information processing in recent years. It enables the structuring of diverse big data using very simple formats without particularizing individual definitions of the data. Several database products are provided that comply with Json. WAGRI adopts Json and a compatible database to realize the adaptive agricultural data structure.

In WAGRI, agricultural data are classified into two categories. One is geospatial data, such as microclimate, soil classification, farmlands. Another one is master data including a dictionary of agricultural terms and master data of agricultural chemicals etc. The following subsections describe the data in detail.

(a) Geospatial data

Vector data of soil types

Most of the data for the cultivation environment are accompanied by minute geospatial attributes, because crop plants are highly sensitive to subtle changes of environmental factors. Figure 23 shows soil data implemented in WAGRI. Headers comprise of information concerning soil categories followed by lots of real numbers with a large number of digits. These numbers represent pairs of latitudes and longitudes. Because the land shape covered by each soil type is extremely complex, approximately 800 pairs of latitudes and longitudes are necessary to express just a few square kilometres of the area. These data are structured as a form of Json. Latitude and longitude pairs are aligned in a clockwise fashion alongside the shape of the area. All of them are nested in the bracket (Figure 23) separated from the headers. Reading pairs in the bracket sequentially and connecting lines between them creates a geospatial graphic, i.e. a polygon, as in Figure 24. As such, all the geospatial data in WAGRI are structured so as to be easily manipulated by software.

Depicting farmland data

Japanese farmlands are characterized by collections of tiny plots individually owned by small-scale farmers, though large-scale farming has been gradually expanding in recent years. Thus, lots of squared polygons representing farmlands appear if you try to depict just a few square kilometres of geospace. This means that depicting just a few square kilometres of farmlands headers

{'PrefectureCode': '5', 'SoilName': 'Gray Lowland Soil', 'SoilLargeCode': 'F2', 'SoilMiddleCode': 'F2', 'SoilSmallCode': 'F2', 'Polygons':
①[{'Coordinates': [{'Latitude': 39.94785327326453, 'Longitude': 139.94917032719306}, {'Latitude': 39.949492549091346, 'Longitude': 139.94972935120717}, {'Latitude': 39.95501187166544, 'Longitude': 139.95837164416596}, {'Latitude': 39.969037792875056, 'Longitude': 139.9794768313014}, {'Latitude': 39.98408389827826, 'Longitude': 140.00188628703512}, {'Latitude': 39.98971239136062, 'Longitude': 140.0109013059738

Figure 23: Soil data in WAGRI

Figure 24: Soil polygons

requires an extremely large number of latitudes and longitudes, which affects rendering performance by the viewers, such as browsers. Json, which originated from JavaScript, the language for interactive websites, has the advantage of automated data reduction in this situation.

Both of “(a)” and “(b)” in Figure 25 are the same farming area, where collections of farmlands’ polygons are layered on soil polygons. In both figures, tiny aligned polygons are those of farmlands whereas complex shaped polygons coloured red and dark yellow are those of soil type. “(a)” represents the zooming out condition and “(b)” represents the zooming in condition. Although the shapes of farmlands are rectangles in general, the shapes around the blue arrow in (a) look to some extent like triangles. Meanwhile, the corresponding polygons are all rectangles. This is the result of automated adaptations of sparsity or density of latitude and longitude data to be rendered. At the condition of zooming out, users’ concerns are capturing an overview of the area, but not looking into each farmland. Therefore it is not crucial for each polygon of farmland to be precisely depicted. However, at the condition of zooming in, looking into each farmland is the users’ concerns, so all the farmland should be depicted in a precise manner. Taking advantage of changing users’ concerns depending on zooming level, JavaScript, referring to Json, dynamically deducts and complements the data for polygons. Because of this dynamic data manipulation, rendering performance gains. This type of dynamic controlling mechanism is widely implemented in geographic information systems including open source software so that writing a few codes realizes such a function.

(a) Zooming Out Landscape
– automated data reduction of farm lands –

(a) Zooming in Landscape
– automated data complement of farm lands –

Figure 25: Polygons of farmlands layered on soil polygons

(b) Master Data

Master data are prepared for two purposes. One is to define domain specific vocabularies. Taking rice cultivars as an example, there are well over 500 officially registered rice cultivars each of which is identified by very minute differences of features comprising approximately 30 attributes. The other vocabularies such as fertilizers and agricultural chemicals comprise minute segmentations requiring very precise definitions.

Apart from defining vocabularies, WAGRI also prepares agricultural thesauruses. Locally segmented knowledge of traditional agriculture has lasted for a long time, resulting in the creation of lots of local jargon. These vocabularies need to be standardized. Also, they can be classified into upper taxa. Table 3 shows the example of ten vocabularies all of which belong to the upper taxon of sowing seeds. Although this table is simplified, the actual thesaurus in WAGRI has nine hierarchical taxa. Information in this thesaurus is not limited to jargon and taxa. It includes information about their usages, such as target of action, place of operation, and targeted crop. As such, the thesaurus of WAGRI consists of a semantic structure enabling knowledge acquisition from a large volume of miscellaneous agricultural texts. An example is given in the next section.

Table 3: An example of an agricultural thesaurus in WAGRI

Category	Name of the Operation	Type of Action	Target of Action	Sub-target of Action	Place of Operation	Example of Crop
Seeding	Seeding	Sowing	Seeds			
Seeding	Direct on Nursery Box	Sowing	Seeds	Nursery Box		Rice Crop
Seeding	Direct Seeding in Flooded Paddy Field	Sowing	Seeds		Paddy Field	Rice Crop
Seeding	Ion Coated Direct Seeding	Sowing	Ion Coated Seeds		Paddy Field	Rice Crop
Seeding	Direct Sowing of Rice on Well-Drained Paddy Field	Sowing	Seeds		Well-Drained Paddy Field	Rice Crop
Seeding	Nursery Bed Seeding	Sowing	Seeds		Nursery Bed Farm Lands	Rice Crop
Seeding	Sowing for Green Manure	Sowing	Seeds for Green Manure			
Seeding	Seeding for Nursery Bed	Sowing		Nursery Bed		
Seeding	Seeding by Planting Machine for Cuttings	Sowing				

WAGRI as a service provider

(a) Services for data retrieval

WAGRI is equipped with a simple user interface to provide users with the services for downloading or manipulating data with good performance. All the interfaces are integrated as a form of Representational State Transfer (REST) API, which is commonly used in Internet services. Figure 26 illustrates the example of requesting a string of API in the case of soil data. '{ }' indicates variable parameters. In this case the parameters are a pair of latitude and longitude. The request string together with information of authentications are sent to the server, then the server sends relevant data back to the client (Figure 27(a)). Figure 27(b) shows an example of code for the request based on Python, just a few codes enables the user to retrieve a large volume of data. Figure 27(b) is also an example of API of soil types, though the request string is different from (a) in the same figure. Figure 27(a) is for requesting 1/200 000 reduction scale of soil area, whereas (b) in the same figure is for requesting 1/50 000 reduction scale. As such, WAGRI prepares diverse APIs that amount to approximately 70 APIs to cope with a variety of users' requests. A designated menu of APIs is prepared on the web site of WAGRI, accompanied by the page of the specification of each API, where simple trial functions are implemented as well.

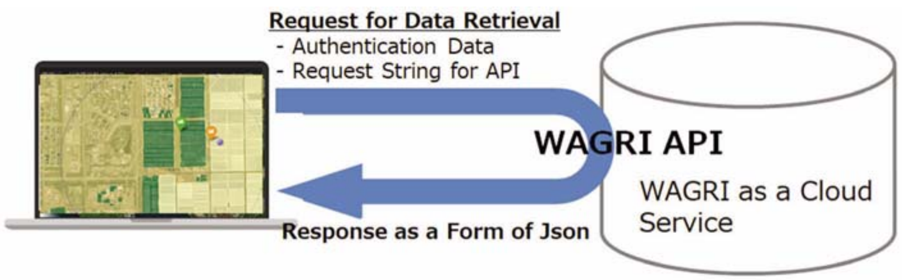
```
API/Public/20_1SoilMap/Get?Latitude={Latitude}&Longitude={Longitude}
```

(a) 1/200,000 Reduction Scale

```
API/Public/5_1SoilMap/Get?Latitude={Latitude}&Longitude={Longitude}
```

(b) 1/50,000 Reduction Scale.

Figure 26: Request string of API in case of soil data



(a) Data Retrieval Based on REST API

```
headers = {authentication information}
Request = urllib.request.Request('https://api.wagri.net/API/
              {parameters for data request}',headers=headers)
response = urllib.request.urlopen(request)
```

(b) An Example of Codes(Python) for Data Retrieaval

Figure 27: Data retrieval from WAGRI

(b) Services for Sensor Data Registration

WAGRI provides not only data retrieval services but also data uploading services. Users can upload their own data to WAGRI and register it. Once users fill in the items of the requested string of APIs for data uploading, the data are sent to WAGRI as post data of HTTP and automatically stored in the form of Json. Among these services, WAGRI pays special attention to sensor data captured from sensors placed in crop fields. Figure 28 illustrates the APIs for sensors. Figure 29(1) is sensor master, controlling general specification of each sensor product such as manufacture, monitoring capability, and precision. Then, users assign an ID to each sensor device and register the information of allocations to their farming fields (Figure 28(2)). Sensors' data cannot be appropriately interpreted without the information obtained by monitoring conditions. To cope with this, monitoring circumstances such as elevations, latitudes and longitudes and the other conditions (Figure 28(3)) can be registered. Finally, data from sensors are stored in a way that each item is identified by a sensor ID (Figure 28(4)).

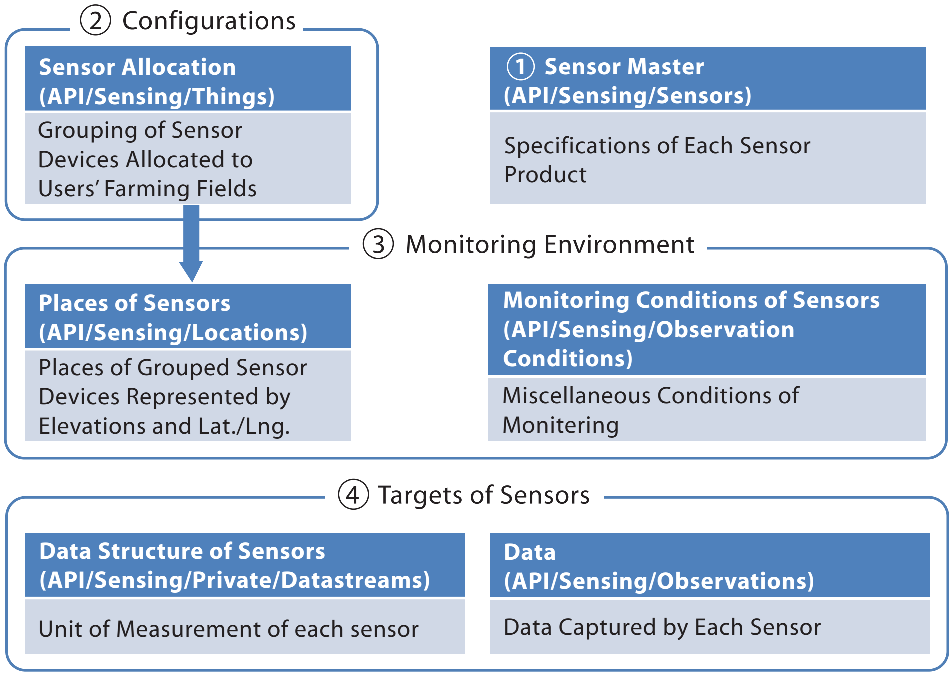


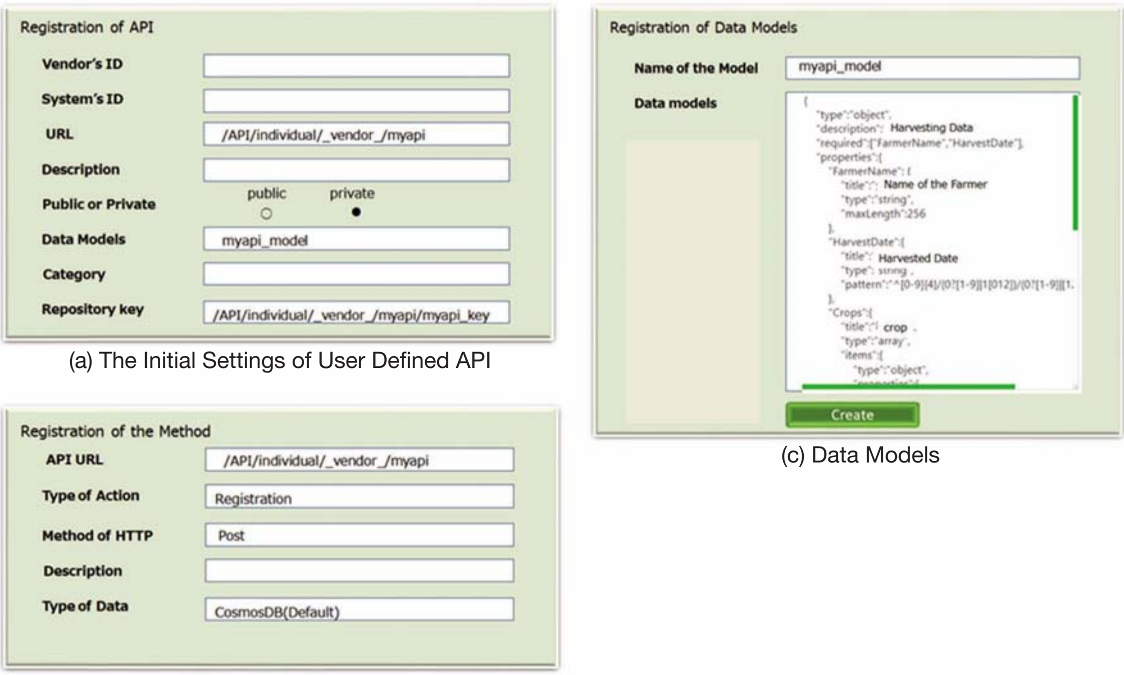
Figure 28: The APIs for sensors

Access control

WAGRI has access control functions to protect agricultural data owned by farmers and private corporations. All the registered data described in the previous subsection are stored on a private domain in the database in WAGRI, strictly partitioned by each data owner. By means of data administration, owners provide restricted users with access privileges. Ownership of data might become complex if secondary uses of the data were to create new knowledge. Although this is not a matter of systems, WAGRI established the data ownership guideline as a reference for forthcoming issues mainly with respect to agricultural data services.

User defined API

Apart from pre-prepared APIs, WAGRI enables users to define their own APIs without any programming, assuming cultivation records, growing conditions such as height of crops, length of leaves, and their colours are observed by farmers and recorded in the form of numerical values with the texts of memos. The items of the records are different depending on each crop and farmer, making it difficult to prepare standardized APIs. In order to cope with this situation, WAGRI is equipped with user defined API. User defined APIs enable users to create their own original APIs without any programming or implementation of databases. Principally it consists of three functions: registering namespace; defining methods; and modeling data structures. Figure 29 illustrates the functions. Figure 29(a) illustrates the initial settings, “URL” is the name of the API and “Repository Key” is the ID of the data in the database that distinguishes it from the other data.



(a) The Initial Settings of User Defined API

(b) Registration of Method

(c) Data Models

Figure 29: Registration of user-defined API

Figure 29 (b) is the registration of the method, which defines how the API manipulates the data. The data structure for the manipulation is defined as a form of Json as in Figure 29(c). By means of Json, any agricultural object can be modeled. Usually, API accompanies the development of programming, implementation of databases, and designing web interfaces taking some time for the construction. However, just these three functions enable complete APIs to be delivered.

Advanced data processing based on dynamic API

In practice, cultivation records are not limited to observational data but the data concerning environments such as microclimate corresponding to the farming fields, and the data captured by sensors. Several issues exist in this case. Updating cycles of these data are different from each other making it difficult to create the records simultaneously. Additionally, the data from sensors are not necessarily stable so that they need to be statistically estimated. For example, observational data are recorded when farmers' daily work finishes. Furthermore, time series data of sensors are stored at constant intervals by the sensor API. Because the precision of sensors are vulnerable to environmental noises, calculating the average etc. might be necessary at the end of daily monitoring. As such, it is not easy to integrate agricultural data as a user oriented structure.

To cope with these kinds of issues, WAGRI prepares a synchronous and/or asynchronous data processing script. Very flexible data processing becomes available by the chained sequence of this script incorporated in the user defined API (Figure 30).

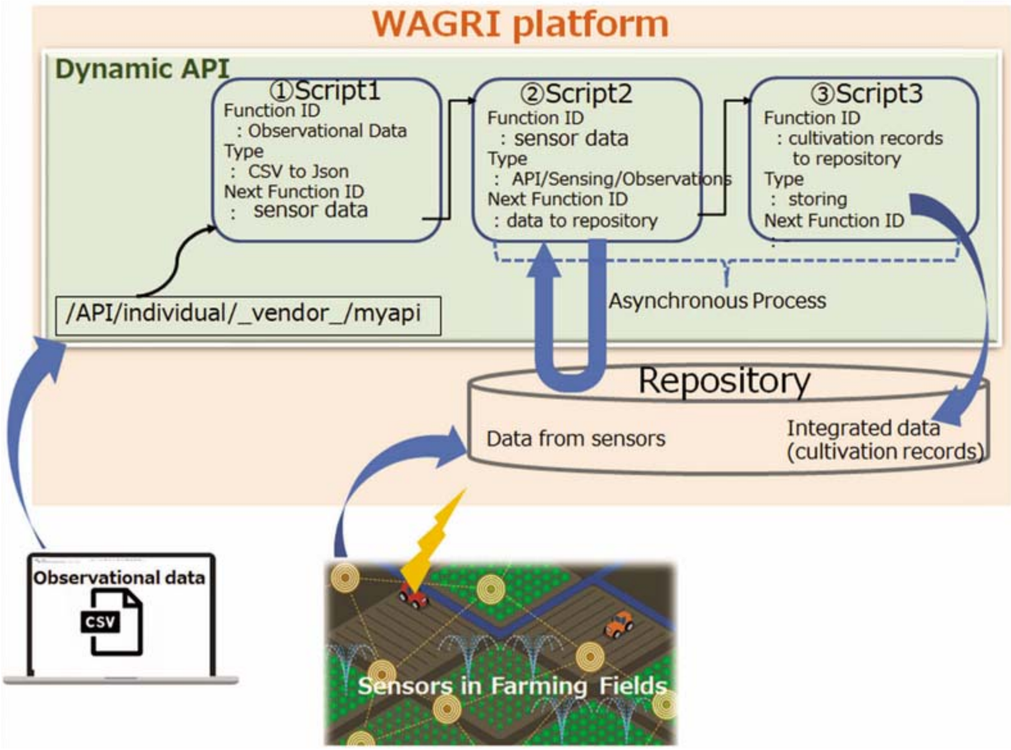


Figure 30: Dynamic API

Once the API is executed, the observational data are uploaded from the client, but not stored in the database in real time. Alternatively, the script is activated and the data are dealt as input to the first script to convert the data format from csv to Json cultivation records, Figure 30(1). Then the script waits for sensor data to be available. The second script internally executes the sensor API to retrieve the sensor data and fill in the relevant items of the cultivation records, Figure 30(2). Finally, the third script stores the integrated data to the database, Figure 30(3). As such, the user defined API combined with the script enables asynchronous, internal execution of multiple data processes in the manner of an organized sequence. This unique architecture of API was developed for WAGRI and is named the “Dynamic API”. In fact, all the APIs provided by WAGRI are based on the architecture of the Dynamic API.

Acquiring knowledge from agricultural big data

(a) Prediction of harvesting

Collecting and storing data are done ultimately to discover useful knowledge, for example for estimating the proper stage of harvesting. In this respect, WAGRI provides users with an algorithm for harvesting prediction. Although the model of harvesting prediction was developed in the 1980s in Japan (Horie and Nakawawa, 1990) it did not prevail in the farming field as described in the introduction. WAGRI enables the model to prevail by realizing it as a form of API.

According to Horie and Nakawawa (1990) the growth of crops is determined by daily temperature and day length. Figure 31 illustrates the relation between temperature and the daily growth rate. This nonlinear curve implies there exists a threshold at T_n such that the growth rate dramatically changes before and after the threshold. Additionally, this threshold together with other parameters depends on cultivar crops. The parameters are so fragmented that the values are varied by rice cultivars, for example. The API internally prepares diverse parameters so as to provide users with the precise predictions. Once users set variables to the API, such as the ID of the cultivar and latitude and longitude of their farming fields, the API selects appropriate parameters corresponding to the selected crop, and refers to the microclimate in WAGRI based on the specified latitude and longitude, then the harvesting prediction is executed.

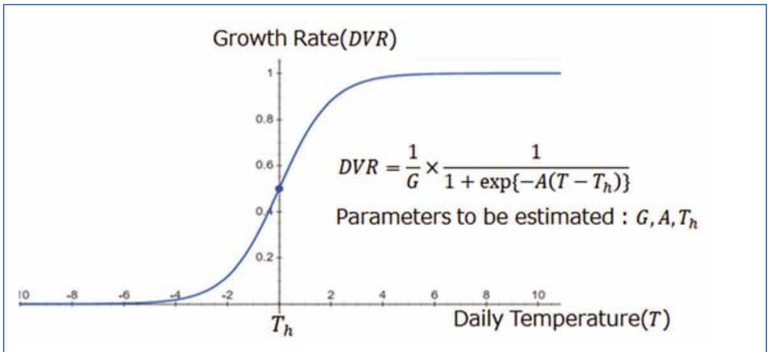


Figure 31: Growth rate based on daily temperature

(b) Discovery of new knowledge

Apart from numerical data, agricultural text data are of great value for knowledge discovery. Indeed, precious know-how might be buried in documents such as farmers' memos, growth records, etc. However, such documents are full of local jargon, which makes semantic analysis by software such as natural language processing difficult. To overcome this issue, use was made of agricultural thesauruses introduced earlier, and miscellaneous agricultural newspaper articles were semantically classified into topics as a trial. We applied latent Dirichlet allocation (Blei *et al*., 2003), an unsupervised machine-learning algorithm, to the documents referring to the thesauruses. Figure 32 illustrates the result making use of the user interface provided by pyLDAvis.[59] Each balloon represents a classified topic within which classified words are included in a semantically consistent manner. Moving the mouse over each balloon show the words by order of their probabilities belonging to the topic as can be seen at the right hand side of Figure 33. Apparently, balloon 2 represents a topic concerning fruit trees. Similarly, other balloons are found to represent other topics including semantically consistent words.

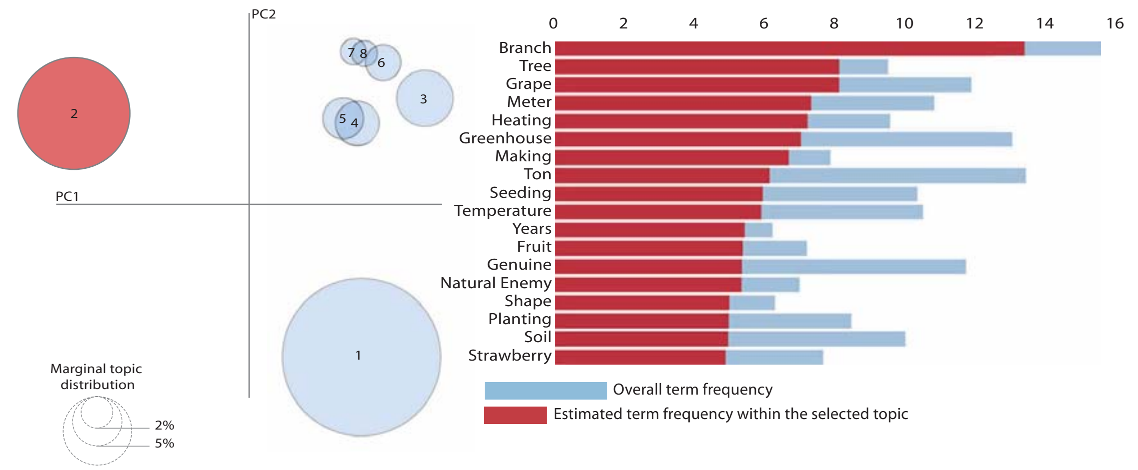


Figure 32: The results of topic model applying to agricultural news articles

Conclusions

WAGRI has been constructed as an unprecedented agricultural big data platform. It is characterized by collecting diverse agricultural data in an integrated manner, and by providing a large number of APIs as a service. Furthermore, Dynamic API, the original architecture of WAGRI, enables users to create their own APIs that realize complex data manipulation using multiple kinds of data. The platform is expected to contribute to the discovery of new agricultural knowledge.

[59] https://github.com/bmabey/pyLDAvis

Acknowledgments

For this article, the valuable support of Mr Kosugi and Mr Sakamaki, NEXTSCAPE Inc., especially with regard to the WAGRI API architecture, is gratefully acknowledged. This work was also supported by the Cabinet Office, Government of Japan, the Cross-ministerial Strategic Innovation Promotion Program (SIP), "Technologies for creating next-generation agriculture, forestry and fisheries". Funding was provided by the Bio-oriented Technology Research Advancement Institution, NARO).

References

Blei, D.M., Ng, A.Y. & Jordan, M.I. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3: 993–1022.

Horie T. & Nakawawa H. 1990. Modelling and prediction of developmental process in Rice: I. Structure and method of parameter estimation of a model for simulating developmental process toward heading. *Japanese Journal of Crop Science*, 59(4): 687–695.

Contact

Hiroshi Uehara (uehara@alcita_pu.ac.jp) and
Atsushi Shinjo (kaminari@sfc.keio.ac.jp)

CGIAR

Platform for
Big Data
in Agriculture

ISBN 978-92-5-131659-7

9 789251 316597

CA5427EN/1/07.19