Sharing more widely: data at the heart of evidence, policy and transformation at Defra

Kath Cameron* and Stefan Janusz*, Vanessa Pilley, Farhana Amin, Steve Wilkinson, Jonathan Hicks, Francesca Gauntlett, Gary Kass, Mike Rose, Sophia Oliver, Emily Miles, Ian Boyd

UK Department for Environment, Food and Rural Affairs
**corresponding authors:*
kathleen.cameron@defra.gsi.gov.uk and
stefan.janusz@defra.gsi.gov.uk

Abstract

Defra (the Department for the Environment, Food and Rural Affairs) and its network of arm's length bodies – the Environment Agency, Natural England, the Marine Management Organisation, Kew Gardens, the Centre for Environment, Fisheries and Aquatic Science and the Animal and Plant Health Agency – produce large amounts of data. This has been instrumental in providing evidence to support the development and delivery of key policies.

Historically, expertise has been embedded in subject-specific areas, resulting in data existing in silos with varying degrees of accessibility for those within the department, wider government and wider society. In 2015, Defra's Secretary of State declared that more than 8000 of Defra's datasets would be made freely available for anyone to access, use and share. Doing so creates opportunities for everyone – not just those making their living in food, farming and the environment. Opening access to Defra's data is intended to both provide opportunities to those who wish to use it to exploit its business potential, and to improve policy delivery by engaging a wider community in solving problems.

The Department has been focused on making datasets available as open data; ensuring future data collection and publication approaches are open by default, taking a transparent approach from the outset. This will make it easier to work more collaboratively with other government bodies, external partners and the public.

Alongside growing support for citizen science activities, new technologies are changing the way in which data flows. This presents opportunities for government as much as for others. Today's smartphones, for example, have the computing power of the supercomputers of 30 years ago. When coupled with a growing range of accurate sensors, smartphones potentially allow the move away from environmental monitoring depending on sparse, fixed, high-accuracy, expensive monitoring stations, replacing them with mobile, cheap, lower accuracy but more densely populated networks, which could be carried on vehicles such as buses or taxis in addition to being supported by the capability of smartphones.

There is already a move towards crowdsourced data for biodiversity measurement (supported by the National Biodiversity Network). Many people, connected together and using mobile phone technology, could revolutionise country-wide data collection that is sufficient to address policy needs. In future, machine learning may allow species recognition by the smartphone in real time when connected to cloud-based software. This kind of completely integrated software and hardware capability is broadly described by the potential of internet-of things-devices. These have the potential to revolutionise the farm: from making comparisons on productivity, to tracing the journey of food from field to fork.

Defra is also explore the use of real-time data from earth observation and remote sensing to respond to incidents such as floods, invasive species, pollution and epidemics affecting livestock. Opportunities form these types of data sources are often balanced by challenges associated with transforming systems to be more data-driven and to ensure that, in areas that are business-critical, the data sources are assured.

Complex decision-making is already supported by modelling that often requires data flows from multiple sources. Machine learning has the potential to increase the automation and speed of data assimilation in to models For example, this could mean using these technologies to inform policies that reward farmers for acting in ways that maximise positive national outcomes. In this particular instance and more widely, however, this comes at the risk of taking human subjectivity and its political and moral weight out of processes.

In future data volumes will be a major challenge. While compression algorithms may improve it is not going to be possible to store all raw data. This is going to require real-time processing of data to provide intermediate outputs. Judgements will need to be made about what data needs to be kept and how to implement intelligent compression focused on the ultimate use.

The main ethical challenges around data concern personal information, including observance of legislation on data protection and more recent legislation on the right to be forgotten. Defra is working both internally, among its network of agencies, and with its data users to ensure these rights are fully respected. However, data ethics extends beyond just consideration of personal data:

there are sensitivities around some datasets for which Defra shares custody, and others where there are implications for civic, national or cyber security. Some data are also commercially sensitive. There may be additional sensitivities where data concerns reporting of environmental monitoring and enforcing regulation.

Funding reforms have also been a major driver for change. Finding innovative ways to save money, but deliver a high quality service, remains a departmental mission – new data policies can help to transform some of the department's ways of working. Adopting an 'open by default' approach has impacts beyond data, as working openly allows more efficient collaborative working that avoids duplication, saves money, and allows cross-pollination of ideas across the Defra group.

Defra is looking more broadly at how to obtain, analyse and manage data (tools, techniques and storage) to further drive innovation. This has been demonstrated by the recently established Earth Observation Centre of Excellence. This centre aims to ensure satellite data is used to its full potential in policy development and operations across Defra by 2020, via a collaborative group of Defra organisations and external partners. A recent success was during the severe weather in December 2015: the centre facilitated better involvement and communication across departments on flood estimates of non-urban areas using radar satellite systems.

Keywords **–** environment data; remote sensing; earth observation; farming;

1 Introduction

Defra (the Department for the Environment, Food and Rural Affairs) is the government department responsible for safeguarding the UK's natural environment, supporting its world-leading food and farming industry, and sustaining a thriving rural economy. Defra's broad remit means it plays a major role in the UK public's day-to-day life, from the food they eat, to the air they breathe, to the water they drink. Defra has an ambitious strategy to create a great place for living, improve services for its customers and increase local delivery of services, taking account of the decision by the British electorate to leave the European Union. Like the whole of government, Defra needs to do this within a reduced budget.

The Defra group, made up of the core Department and 33 arm's-length bodies, produces vast amounts of data, providing a robust evidence base for key policies and monitoring their delivery. Defra recognises the importance of its data. The opportunities better use of data offers to improve policy-making and delivery – providing a better service to Defra's customer while also saving money – have all contributed to a growing interest in the data agenda.

A revolution in data has come about, thanks to the pervasiveness of data itself: mobile devices run apps allowing people to have a vast array of data recording and analysis tools at their fingertips; there are new and cheaper ways of sensing, collecting and exploiting data. These provide the technological foundations for a modern, open and data-driven department. While all these opportunities exist, Defra will also have to ensure it is prepared to exploit them with the right technologies, tools, skills and processes, which comply with legislative requirements and safeguard people's privacy.

2 What have we done so far?

2.1 Releasing Defra's data

While some of the vast quantities of data that Defra holds was in the public domain (for example, imagery from the Copernicus satellite system; real-time air quality and river level data; National Biodiversity Network records) much has historically been held internally, and has not been easily accessible outside Defra or the arms-length bodies that acquired it. Using data from one part of the organisation in another was equally problematic: the data was siloed, often covered by restrictive licensing conditions, and in many cases non-interoperable. Data was duplicated and records in duplicated datasets sometimes did not match.

In June 2015, Defra's then-Secretary of State set a target of releasing 8000 of Defra's datasets within a year. Doing so brings a number of immediate benefits, largely resulting from reduced friction in access to useful datasets. Licensing the data as 'open' (under an Open Government Licence) enables anyone to access, use and share it, including for commerical innovations in the private sector, which creates economic growth and reduces the burden on government to provide innovative services with the data – something neither the government nor in fact any organisation has a particularly strong track record in. McKinsey (2013) [1] estimated that, globally, $3 trillion could be unlocked by opening up such datasets. Releasing the data under an OGL Licence also makes it easier to use, reuse and share data across the Defra group and across wider government, reducing friction and improving efficiency; while the not-for-profit and research sectors, are able to use, analyse and arrive at solutions provided by the data much more readily.

The target chosen was largely arbitrary, although based on a 'good guess' of how much non-personal data Defra held. It is large enough to require a concerted effort from across the group, and in fact the effort was focused

between 10 areas of Defra, including several of the larger ALBs including Environment Agency, Natural England, the Rural Payments Agency and the Animal and Plant Health Agency. However, it wasn't so large as to be unattainable, or lead to deliberate 'salami-slicing' of the data to meet the target.

A data programme team was assembled and tasked with a number of workstreams, to ensure the target was met but also to drive the cultural transformation, systems architecture design and skills development to deliver a data driven department. In the first phase, which finished in June 2016, this included leading on, facilitating and assisting teams across Defra group to publish open datasets. The team first had to identify what data was held and what could be released, and then develop publication processes, including a common light-touch risk assessment to ensure releases did not breach data protection or information rights legislation. Despite the ad hoc approach to building the capacity to deliver the target, over 10,000 datasets were released within the year. Now, over 12,000 open datasets have been published as of mid-August 2016, meaning that 42% of open government data with a record on data.gov.uk comes from Defra. Datasets released have covered the whole range of Defra's activities: family food surveys stretching back to the war; data on bee colony health; on land and marine protected areas; biodiversity; and flood maps and modelling – to name a few.

In some instances, data to be released had previously been licensed commercially, so the decision to release the data openly also had to take account of loss of income. An example of such data is Light Detection and Ranging (LiDAR) data. For the last 17 years, the Environment Agency has used LiDAR to scan and map the landscape from above, collecting dense and accurate elevation data which is used in a number of ways: creating flood models; assessing coastal change; and analysing land use. In September 2015, the Environment Agency made all of its LiDAR data (approximately 11TB in size) available as open data. This has resulted in unanticipated uses, such as:

- archaeology – visualising ancient landscapes to find historical features such as Roman roads; [2]
- computer games – importing into virtual worlds in *Minecraft* and *Open Transport Tycoon*; [3]
- Orienteering – to develop slope maps and walking apps; [4]
- Civil engineering – to manage infrastructure and plan projects, including solar cell installation and planning watercourses [5]

There are also datasets that tell particularly engaging stories. One example is the Family Food Survey: records from food diaries completed by 150,000 households between 1974–2000 was released as open data in February 2016 after some changes to protect the privacy of those who had responded to the surveys. [6] This data shows how the typical British diet has moved away from white bread, whole milk and eggs to more brown bread, skimmed milk, cereals, fewer eggs and more exotic vegetables. The data is of interest to health officials, historians, food retailers and white goods manufacturers, who can see how trends follow a combination of advances in technology, public health campaigns and changing social norms. Another example is the single use carrier bag use data, which was released in early July and indicated an 85% reduction in carrier bag use since the introduction of the 5p charge in England. [7]

At the end of August 2016, Defra's open datasets had been accessed 200482 times and downloaded 53771 times from data.gov.uk.

2.2 Plans for further data releases

Defra still holds a number of datasets that have not yet been released as open data. For some of this data there are questions about whether the data is personal, or there are particular sensitivities which may prevent or restrict what can be released. In making decisions about releases we will need to understand the benefits of being more open against the risks, as well as developing and implementing processes and agreements to enable release with the appropriate safeguards. Plans are being developed to identify and address the issues to enable further releases.

Some future releases will update existing datasets. This may include improving the quality or usability of the data (integrating user feedback) and improving the quality of metadata, including implementing standard tagging and name conventions.

There are a number of complex ethical challenges around data that need to be considered. Commonly these challenges surround personal information, including observance of legislation on data protection and more recent legislation on the right to be forgotten. Defra is working both internally, among its network of agencies, and with its data users to ensure these rights are fully respected. However, data ethics extends beyond just consideration of personal data: there are sensitivities around some datasets for which Defra shares custody, and others where there are implications for civic, national or cyber security.

2.3 Becoming data-driven

Defra wants to become a modern, open and data-driven organisation, using data and its own data resources in the best way possible to achieve its objectives. Historically, the Defra group has collected and managed data based on the needs of individual organisations and teams. There is no coherent data management across the group, and identifying custodians of datasets that span areas is is not always trivial. Many of the existing systems and processes

do not have the flexibility or capacity to take advantage of new ways of working and new opportunities.

In addition to taking an 'open-by-design' approach, which reduces friction by making it easier to access and use data across Defra, there is a simultaneous cultural change which has begun to treat data as a valued asset, with a more coherent approach to how data is acquired, managed and shared. It is believed that doing so will enable better use of internal and external data to develop strategies, make decisions and more respond quickly and effectively during emergencies (as has already happened during flood incidents in December 2015). Achieving the goal brings a number of challenges, which we are working to address:

- A *central data platform* to enable internal and external sharing of data. This is linked to Defra's IT transformation replacing existing ICT services, and will enable closer and more coordinated working. Options for the platform are being explored
- Mechanisms to enable Defra data to be easily *discovered* (e.g. comprehensible metadata) *and accessed* (eg APIs and Apps). We anticipate that some APIs and apps may be developed within Defra where the Department offers a particular expertise, but we will be looking the market to develop and innovate with our open data
- *Managing data to a common framework* including a common lifecycle. More coherence and consistency across the Defra group will enable common approaches to data acquisition, management, standards (e.g. for metadata, customer records, licensing etc.), sharing and archiving and closer working together.
- Increasing general *data literacy and skills* across the group and ensuring data experts can access training and the latest thinking on new analytical tools and techniques.
- Understanding *user needs,* both internally and externally, to make the best use of our data. This is potentially an enormous task and Defra will need to prioritise attention on a number of key business areas to realise real benefits.
- Understanding and exploiting *new opportunities offered by data science*. Defra's Science Advisory Council, advising the Chief Scientific Advisor, has established a subgroup focusing on data science opportunities and risks and Defra is also exploring establishing a Data Science Fellowship with the Science and Technology Facilities Council.

Plans are being developed to achieve all the above. The Data Programme is working with Defra's network of agencies, its IT and digital Transformation Programmes, the Government Digital Service and other Departments on cross-Government requirements and opportunities.

3 Future opportunities

3.1 Earth Observation

Defra has established the Earth Observation Centre of Excellence (EOCoE) with external partners, including the UK Space Agency, to explore the potential of satellite data and other types of earth observation data in helping to deliver policy and services. Data types being explored include optical, thermal, radar and LiDAR as well as Sentinel data from the EU Copernicus programme and NASA/USGS Landsat data. The EOCoE programme is expected to deliver significant benefits, including reductions in field inspections and more effective delivery of a range of Defra group functions including forestry management and land management, and flood incident response.

The Centre of Excellence works with policy teams on potential opportunities, and with external stakeholders to explore services of value to Defra, in addition to managing pilot projects exploring new applications of EO data.

Pilot projects exploring new applications include:

- *Forest monitoring* where measures of the amount of vegetation obtained through multispectral satellite imagery are matched to National Forestry Inventory, GIS, and aerial data. This provides a more complete picture of rates of harvesting and restocking, giving clearer evidence for the impact of industry investment, carbon accounting and natural capital.
- *Water quality monitoring* exploring where use of earth observation data e.g. crop cover, land use etc. – combined with other data, e.g. on livestock, can help to target monitoring. This would save money while maintaining capability to detect changes in water quality and support actions to address these.
- *Habitat condition and ecosystem services* exploring where earth observation data products can be used to streamline monitoring of protected grassland habitats, reducing monitoring costs and enabling targeting of field staff to high risk changes affecting biodiversity or ecosystem services. It will also enable monitoring of the impacts of interventions for example through Countryside Stewardship Schemes.
- *Farm inspections* exploring whether earth observation data combined with new analytical techniques can detect land use change.
- *Habitat management* exploring whether earth observation imagery can be used to assess habitat condition reducing the need for repeated and costly field surveys.
- *Living maps for biodiversity and natural capital* exploring the creation of maps using novel processing methods, combining satellite and ground survey data, enabling landscape change detection and targeting efforts

to manage biodiversity. The Norfolk and Northumberland National Park living Maps have allowed better monitoring and control of land erosion, and made it possible to plan green infrastructure, including managing land use in accordance with the national pollinators' strategy (e.g. *Bees' Needs*).

- *An internal collaboration node* enabling sharing of geographic information including earth observation data across the Defra group, partners and contractors which is expected to merge with the central data platform described above.

Earth observation data is used increasingly as an information source for disaster management and emergency response. In the UK, earth observation is successful for monitoring flood extent in rural areas. Imagery was used extensively during the flooding of Somerset levels in early 2014 and more recently during the December 2015 flooding in Cumbria, Lancashire and Yorkshire. In the latter incident, satellite data was used to produce flood estimates of rural and non-rural areas, supporting communications and management of recovery work. This includes assisting the Rural Payments Agency to determine affected farm holdings and target assistance payments.

There is potential for Earth observation through all parts of the flood emergency cycle:

- *Strategic flood risk planning*. Airborne Earth Observation data (from aircraft) already underpins the strategic flood risk planning of the EA, through the use of airborne LIDAR data for flood mapping. Recognising and assessing the impact of changes to land use within catchments, eg. additional hard-standing defences; tree cover; ditches; barriers; roads, etc. – on an annual basis, through access to very high spatial accuracy data would further support planning.
- *Flood Defence Asset Inspection*. Many assets are extensive earth banks which are inspected for their integrity. Remote change-identification methods using hi-res imagery and SAR could offer a potentially cost-effective partial solution when used in combination with airborne and ground measurements. This is still to be tested.
- *Emergency flood mapping* to monitor flood extent. Over large areas such as the Somerset Levels satellite data is currently the only viable monitoring technology and is now used routinely in such situations.
- *Flood Recovery* requires a considerable amount of effort. Currently, aerial photography and airborne LIDAR (pre- and post-flood) are widely used as a basic planning and management tool – particularly for extensive tidal inundation recovery. Very high resolution annual data could provide much more current information on pre-flood status than is available at present, and would allow targeted additional acquisition of comparable data post flood by the Environment Agency.

To date, there has been limited use of earth observation for other problems such as invasive species, pollution incidents and epidemics affecting livestock. Often, earth observation is used in combination with other data and mathematical models to assess risks, likelihood of spread of diseases, etc.

A significant challenge to earth observation is the volume of data collected. Data could be filtered and/or compressed before storage, but in addition to removing noise, it may be that data that can be usefully explored to provide insight to planetary-wide change in the future is discarded because means to extract and use it are limited at present. The velocity, variety and intercompatibility of earth observation datasets also presents a challenge, in terms of interoperability and making sense of the complexity of data being presented.

3.2 Sensors

Historically, Defra has used more traditional ways of gathering data – for example, through field sampling and surveys. With the increasing availability of low-cost sensors, there is enormous potential to replace traditional monitoring approaches with automated ones. This will mean lower running costs, but an increased need to handle the large volumes of data resulting from constant monitoring. Defra already collects a range of data via sensors which automatically record and transmit data:

- Automatic Urban and Rural Network – UK's largest automatic network of monitoring stations providing hourly information on a range of air pollutants. Used for reporting against EU Ambient Air Quality Directives and providing information to the public
- Flood Information Service River and sea levels in England are measured twice daily (more frequently during flooding incidents) from monitoring stations across the country; they form the basis for flood response
- SmartBuoys – autonomous, moored systems providing high frequency data on marine environmental quality

Much of this sensor data is available publicly along with tools to make it easily accessible to the general public. Defra is exploring other opportunities for collecting data remotely. These include the use of autonomous underwater vehicles to collect a range of marine environmental information without the significant costs associated with ship time.

In the future, data collection, in addition to use could be democrastised. Alongside growing support for citizen science activities, new technologies afford revolutionary possibilities. Today's smartphones, for example, have the computing power of the supercomputers of 30 years ago,

coupled with a growing range of accurate sensors that improve and are added to with every new generation. Smartphones potentially allow the move away from environmental monitoring depending on sparse, fixed, high-accuracy, expensive monitoring stations to dense, mobile, low-accuracy, cheap networks many of which could be carried by people or placed as remote sensing devices on vehicles such as buses or taxis.

There could be a move towards crowdsourcing data collection on biodiversity, where many pairs of eyes, connected to mobile phone technology, could revolutionise country-wide data collection that is sufficient to address policy needs. Internet-of things-devices have the potential to revolutionise the farm: from making comparisons on productivity, to tracing the journey of food from field to fork.

3.3 New approaches

Defra is just starting to explore the opportunities offered by new approaches to understanding some of our larger and more complex datasets. One area is in machine learning, where there is an interest in harnessing the technology to effectively predict, detect and respond to hazards and threats. In conjunction with the increasing capabilities of machine learning, complex decision-making which requires integration across multiple sources could be made more quickly and with greater objectivity; for Defra, this could, for instance, mean using such technologies to inform policies that reward farmers for acting in ways that maximise positive national outcomes. In this particular instance and more widely, however, this comes at the risk of taking human subjectivity and its political and moral weight out of processes. Human judgement will very likely still be critical in in making the most important decisions based on data, using machine learning's capacity to determine the critical factors in complex systems to augment human insight.

Novel analysis techniques are also expected to reduce the costs of inspections and increase regulatory compliance. Using the large amount of data Defra holds about farms, as well as data on external factors it should be possible to identify those farms that pose the biggest risk of being in breech and target inspections towards these reducing costs for government as well as for farmers. Machine learning techniques could be used to explore a whole range of data sources against farm inspection history to spot which data values are most closely associated with failing an inspection. Knowing that should refine predictions of which farms to visit and also the risk level of probability of missing farms that should be inspected.

Conclusion

Data science and its importance to evidence is now acknowledged as being fundamental to strategic planning across Defra group. Data is at the heart of ongoing transformational programmes intended to deliver better services to customers at lower cost. It is recognised that building capacity towards a more data-literate workforce within the groups will allow Defra to make the leap between Defra's wealth of datasets and the crucial information they contain to allow ever-more effective development and delivery of policies.

There is also a recognition that expertise exists outside of the group as well as inside, and that explicitly taking an open-by-design approach to data is one way to capitalise on expertise and innovation in the private, not-for-profit and research sectors as well as wider government. Defra's desire is to become an organisation that has an open approach to data and drives forward this agenda across Whitehall. This requires radically changing our approach to data and how it is used, from the bottom-up, to the top-down.

Acknowledgements

We thank everyone in the Defra Data Programme team, Earth Observation Centre of Excellence, and Defra Chief Scientist's office who had input into this paper.

References

[1] McKinsey Global Institute. 2013. Open data: Unlocking innovation and performance with liquid information

[2] Lasers reveal lost Roman Roads (3 Feb 2016) https://www.gov.uk/government/news/lasers-reveal-lost-roman-roads

[3,4] Environment Agency: Laser surveys light up open data (18 September 2015) https://environmentagency.blog.gov.uk/2015/09/18/laser-surveys-light-up-open-data/

[5] Exploring the Laser 'Cloud' (17 March 2016) https://defradigital.blog.gov.uk/2016/03/17/exploring-the-laser-cloud/

[6] Family Food Open Data (18 February 2016) https://www.gov.uk/government/statistics/family-food-open-data

[7] https://www.gov.uk/government/publications/carrier-bag-charge-summary-of-data-in-england-for-2015-to-2016