Defra Science Advisory Council: Data sub group

Report: How can Defra become a fully data driven department?

March 2018

Contents

Introduction

The Defra Science Advisory Council (SAC) initiated a Data sub group to explore and provide high-level advice on the question "How can Defra become a fully data driven department?" by considering what Defra should be aware of and prepare for.

The background to the review includes:

- Vast data reserves from Defra are set to transform the world of food and farming, our understanding of the environment, and our response to events such as flooding.
- Virtually all the data Defra holds will be made freely available to the public, putting Britain at the forefront of the data revolution. Over 12,000 datasets have been released so far.
- Open data will help the UK to achieve its full potential and facilitate becoming a one-nation economy, where the productivity of the countryside will be brought up to the level of our towns and cities.
- In April 2017 Defra published a set of nine data principles to guide all work with data[1].

Defra has challenged itself around questions about how best to exploit readily available data sets to address some of the major challenges it faces. For example, the data collection required to underpin Britain's exit from the EU or the 25 year environment plan. The SAC noted that Defra is not alone in these data-led challenges but sits alongside cross-government and national capability. At the end of the process the sub group considered that the initial challenge could be recast as "Is Defra ready for the challenge of open policy making informed by open-data?"

The terms of reference are at **Annex A**, however, in summary the group set out consider 4 questions:

1. Where is data science going and what should Defra be aware of and prepare for?
2. How resilient is Defra?
3. How should Defra address data quality issues?
4. What are the unintended consequences of releasing/sharing data (data ethics)?

The sub group ran two workshops to address the terms of reference. The first considered the data landscape and mapping to look at question of where is data science going and what should Defra be aware of and prepare for. The second workshop focussed on data ethics and examined the potential unintended consequences of releasing/sharing data. Questions of how resilient Defra is and data quality issues were discussed in both of workshops and the sub group makes recommendations relating to all the terms of reference at the end of this report.

The sub group invited external experts to contribute to the discussions of both workshops (listed in **Annex B**) as well as relevant Defra policy officials to ensure informed and

[1] https://defradigital.blog.gov.uk/2017/04/04/defras-data-principles/

relevant discussions. Inviting Defra officials also enabled workshop findings to be immediately considered and acted on by Defra, in particular the Data Programme team.

In preparing this summary report, the discussions from the workshops have been combined with the work of a STFC[2] fellow (Dr Caroline Poulsen) seconded in to Defra for a year to provide advice on data issues.

Landscape and challenges identified

Data is pervasive, in the Defra context it encompasses the natural and physical sciences to the quantitative social sciences. A challenge for Defra is to move away from 'silos' of data specific to a policy area, e.g. the environment, farming or air quality, and to move towards a data capability that enables a more integrated, cross topic, even holistic policy making approach. For example, following EU exit agri–environment schemes will require modelling and data relevant to both farming and the natural environment, similarly air quality impacts both people and natural habitats. Sharing of data is often restricted by licencing arrangements and a risk adverse approach to sharing potentially what can be seen as "sensitive" data. A possible solution would be the creation of 'safe havens' for people to access and work with sensitive data. This concept of sharing data across policy areas and externally is linked to several of the recommendations the sub group make later in the report.

Furthermore, the way that Defra collects and uses data is changing. Defra currently collects a large variety of data sets ranging from data on the state of the oceans, to information about farms and the state of the environment. Defra also runs a number of operational services from air quality to waste permits. In parallel a new range of external 'big data' sources are becoming increasingly accessible, ranging from social media data such as Facebook and twitter to mobile phone data, earth observation and the internet of things. The different data sets are enabling Defra to use data in new ways such as in the field of artificial intelligence and machine learning. The combination of new and old data sets and analytics has the potential to change the way we analyse data to save resource and drive productivity. However the current infrastructure to enable access and use of the data is not fit for purpose.

To address these changes it will be necessary for Defra to enable upskilling of the work force. There is overwhelming consensus in Defra that the way data is analysed is changing and new skills such as coding to extract, manipulate and visualise data are now a basic tool kit for many jobs. There is a strong demand from within Defra for programming skills, as evidenced by a recent oversubscribed Code Skills Incubator Initiative. The Initiative was set up at working level and involved over 60 Defra analysts developing data science knowledge. To match these skills it is necessary for Defra to provide fit for purpose infrastructure for increasingly complex analysis and which matches the

[2] Science and Technology Facilities Council

requirements of analysts such as easily accessible cloud space, archive and software tools.

The sub group identified four clear challenges that could be important drivers in the Defra data landscape, and these require greater consideration by Defra:

1. The paradox of the data society. The more data we get potentially the less we know about them.
2. Quantifying the value of data. What is the value of making data freely available and how do you quantify this? The economic benefits of improved data access, re-use and inter-operability measured against the cost and resources to implement best practice for open data.
3. Privacy issues. This includes transparency, openness and fairness of decision making, and responsibly using data within an agreed ethical framework (more on ethics below).
4. Responsibly using data within an agreed ethical framework.

Data ethics

Data ethics is extremely important to consider in the Defra context. The data-sets that Defra collects are extremely valuable and have potential use for many sectors including farmers, regulators and commodity traders. Many of the data-sets that Defra holds contain personal information. Advanced data analytics is making it easier to combine data from different sources to make useful insights and innovative products. However, this same technology is making it easier to link data and identify individuals and behaviour. Defra's responsibilities with regards these data sets is not clear but several guidance documents have been produced[3] and provide a starting point for Defra to develop this work further.

Data ethics is a rapidly evolving and challenging field where there are few binary choices and context remains critical. Data linkage and flow-down of consent for example has a degree of ethical complexity that is not currently fully recognised.

The challenge of ethics brings together technical, legal and organisational issues, as shown in the diagram below. On a simple level, the Defra data principles could be extended to encompass an ethical dimension, e.g. data sets should have a meta-data requirement to indicate how sensitive the data is and the restrictions for distribution and archiving. The meta- information/consent needs to flow with the dataset and the uncertainty on the data should be communicated with the data set to combat misuse.

The consequences of ignoring ethics are likely to be a substantial detriment to Defra. For example, will the full potential of farm-based datasets be realised without an ethical framework being constructed.

[3] For example the Cabinet Office Data Science Ethical Framework (2016) or the Royal Society Data Governance report (June 2017).

The top line assessment of the sub group is that Defra, among many other organisations, is currently under-developed in its understanding and practice in respect to data ethics. In Defra data sharing can be even more complex because of sharing restrictions within the Defra group.

Overall observations and recommendations

There is much to be recognised and credited in Defra's work on and with data to date. In particular in making data sets open and providing guidance in the form of the 'data principles' in order to encourage better stewardship of the data.

However, the SAC encourages Defra and the Data Program to encompass more than "making data open" to begin to address behaviour and culture change across the organisation including senior engagement, data science skills, improved data quality (which is a key blocker to enabling wider use), improved documentation and accessibility as well as support and infrastructure for advanced analytics and visualisation.

To address these challenges the SAC makes following recommendations

1. **The SAC recommend the formation of a Defra Data Science Hub to act as a focal point for data science, develop a community of data science users across the Defra Group and communicate and disseminate relevant information**. This community could deliver significant benefits and build on current initiatives underway in the department. The hub could facilitate the sharing of data across the Defra group and move away from 'silos' of data. The hub would enable a good understanding of the data science practitioners and their skills across Defra and in this way the hub could also be used as a way for Defra to actively support upskilling the workforce (recommendation 2). The hub could also provide the essential user input and feedback necessary to develop key analytical infrastructure and support (see recommendation 3 below).

2. **The SAC recommend that Defra put significant resource into upskilling the work force to enable them to apply new data science techniques that can increase productivity and deliver new insights in operations and policy as well as understand an ethical implications**. Defra should build on the momentum developed by the Code Skills Incubator Initiative as well as structured training opportunities and courses to ensure senior officials have a good understanding of data science. In addition it could include training on data ethics (for example an online course similar to that currently offered on handling of sensitive documents), this would provide clarity around what data can be shared from an ethical perspective in order to encourage the use of sensitive data sets within Defra and also to combat any potential misuse.

3. **The SAC recommend that Defra improve the infrastructure (catalogue, computing, software, archive) to enable the increasingly complex cross department analysis that provides the evidence to feed into policy development and enables operational activities**. Defra employees find it almost impossible to identify which data sets exist in Defra. This could be improved through developing and

promoting a data catalogue. The tools, e.g. cloud space and software to analyse the data sets is difficult to obtain. Processes need to be in place and visible to ease access to the tools that enable even simple analysis let alone more advanced analysis such as AI which could be truly transformational.

4. **The SAC recommend that Defra consider the creation of an independent Data Ombudsman to provide a suitable authority for data ethics and usage with both internal and external remit.** The data ombudsman should be safe space to discuss ethical issues in a friendly environment giving guidance for Defra scientists generating their own data sets on what can and cannot be shared and disseminated. In forming decisions on data ethics Defra should use the Cabinet Office (May 2016) Data Science Ethical Framework as a guide. The ombudsman would be available to external users of Defra and Defra-supported data to facilitate fair usage and support consideration of ethical issues.

5. **The SAC recommend Defra explore whether adoption of open source algorithms/practice across Defra should be mandatory.** For example, should the requirement for open source algorithms be mandatory when new research and operational contracts are awarded. This would allow greater flexibility, future proofing, as part of a community of users and potential anti-competitive contractor practice.

6. **The SAC recommend Defra increase the links with expert communities, external to Defra, to foster increased uptake in research and commercial activities as well as expertise in the stewardship of data sets.** The use of Defra's open data outside of Defra has so far been low, linking with external communities, whether they be academic or at a European or International level can often be of benefit in providing increased insights into the data or in-kind resource benefit. A good example of this is the Earth Observation Centre of Excellence which has engaged with the academic community through running short Proof of Concept studies and nurtured a stake in the academic archive and cloud CEDA (Centre for Environmental Data Analysis).

There remains a question, beyond the scope of Defra SAC but which should be addressed by Defra, as to how much resource Defra should expend making data open and accessible (internally/externally).

Annex A: Terms of Reference: SAC-Data Sub Group (October 2016)

Overarching aim

The overarching aim of the SAC-Data sub group is to describe 'How Defra can become a fully data driven department'.

In doing so, the Group will look to learn from work led by others in this area (for example, from across government and from the private sector[4]), but they will also provide an external (independent) perspective, and offer advice and recommendations on the specific issues Defra is, or will be facing.

Background

- Vast data reserves from Defra are set to transform the world of food and farming, our understanding of the environment, and our response to events such as flooding.
- Virtually all the data Defra holds will be made freely available to the public, putting Britain at the forefront of the data revolution. Over 12,000 datasets have been released so far.
- Open data will help the UK to achieve its full potential and facilitate becoming a one-nation economy, where the productivity of the countryside will be brought up to the level of our towns and cities.

Purpose of review

The sub group remit will be broad, and include issues around data science, capability, data quality and ethics. The group will focus on and address a specific set of questions aimed to assist Defra in the development of its Data programme. It will make recommendations, via a report, on 'How Defra can become a fully data driven department':

Specific questions:

1. **Where is data science[5] going and what should Defra be aware of and prepare for?**

[4] The group may want to consider:
- What policies do other research funders have for data?
- What can Defra learn from other organisations about data management and how should it recognise the role of the private sector (potentially a major player)?
- What can Defra emulate and what should it avoid?
- Is there an opportunity to work with device manufactures?

[5] Defra is working to recruit a Science and Technologies Facilities Council fellow to work with the Chief Scientific Adviser's office and the Defra data programme on data science and policy. It is also looking to second a Defra employee into the Alan Turing Institute to work on a specific Defra data project.

- How can Defra best exploit its readily available data sets to address some of the major challenges it faces – for example, the data collection and monitoring required to underpin the EU exit strategy, or the 25 year environment plan.
- Where should Defra data planning sit alongside cross government and national capability plans?

2. How resilient is Defra?

- What is Defra's internal capability in terms of data science/tools/apps? Will this capability adequately support the scale of challenge the department faces?
- Is the department prepared for fundamental changes in its relationship with Stakeholders, and what might fundamentally challenge the way in which Defra works with changes to the way data is acquired, managed and shared?

3. How should Defra address data quality issues?

- How is the quality of data measured, and how can Defra assess this in terms of its suitability for wider use? How is poor data managed?
- Are there lessons to be learnt by virtue of the way data is collected, and is the reach of the data within the expectations of partners?
- To what degree does public data stimulate decision making by adding value and distinguishing between data acquisitions and making data open?
- Are there exemplars, which could be showcased, to demonstrate how the data initiative has changed Defra policy development?
- What mechanisms are in place to ensure, if Defra data is used by a third party, that Defra receives recognition/appropriate citation?

4. What are the unintended consequences of releasing/sharing data (Data ethics)?
For example, risks of breaching legislation, of releasing personal information, copyright issues, privacy issues or the risk to Defra's reputation.

- What are the unforeseen opportunities and risks (e.g. presented through horizon scanning), such as the open data enabling citizens to interpret, design and contribute to policy-making.
- What are the issues for Defra when faced with using data that is not entirely open – e.g. data from part privatised companies etc. that benefitted from Defra sponsorship.

There are other areas of ongoing Defra work where the data sub-group's opinion would be valued. We anticipate engaging the data sub-group in these areas as thinking develops. Issues could include:

- How can Defra enhance its understanding of the link between data provision and data use, and how different people in the organisation (policy makers, analysts, delivery experts), interact with the increasingly available data resources.
- What could enable Defra to establish and undertake new data concepts, including developing additional tools to manage complexities?
- How should Defra utilise its capability to ensure the Defra group is more joined up across its network?
- General advice to the data programme on prioritisation and communication.

Membership

The group will be chaired by Professor Paul Monks and will include Defra SAC members Professor Charles Godfray, Professor Sheila Bird and Professor Wayne Powell. A small number of independent academic co-optees will provide additional expertise. Co-optees are not recruited through open competition, but are appointed based on their specific skills and experience. The SAC-Data co-optees act independently of any of their other interests.

The SAC-Data sub group will be supported by Defra Officials including the Head of Defra's Environment Analysis Unit, Head of Data Engagement, Data Programme and OpenData communications.

Outputs

A review report with recommendations to Defra's Chief Scientific Advisor (CSA) will be presented to Defra SAC.

Duration of sub group

The expectation is that this sub group will be closed once it has addressed the specific questions being posed by CSA/Defra. The group will be set up initially for one year at which point it will be reviewed.

Annex B: External experts involved in the sub group workshops

Data landscape mapping workshop

- Timandra Harkness, comedienne, science writer and broadcaster
- Dr Sabine Hauert, Lecturer in Robotics, Department of Engineering Mathematics, University of Bristol
- Professor Valerie Isham, Professor of Probability and Statistics, University College London
- Professor Neil Lawrence, Professor of Machine Learning and Computational Biology, University of Sheffield

Data ethics workshop

- Daniel Chase, Digital Ethics Lab, Oxford Internet Institute, University of Oxford
- Dr Murray Gardner, University of Oxford
- Professor Ruth Gilbert, Professor of Clinical Epidemiology, University College London Institute of Child Health
- Professor M C Schraefel, Computer Science and Human Performance, University of Southampton
- Professor Richard Tiffin, Director of Science, Agrimetrics, University of Reading

This publication is available at www.gov.uk/government/publications

Any enquiries regarding this publication should be sent to us at

science.advisory.council@defra.gsi.gov.uk