UKEOF
ENVIRONMENTAL OBSERVATION FRAMEWORK

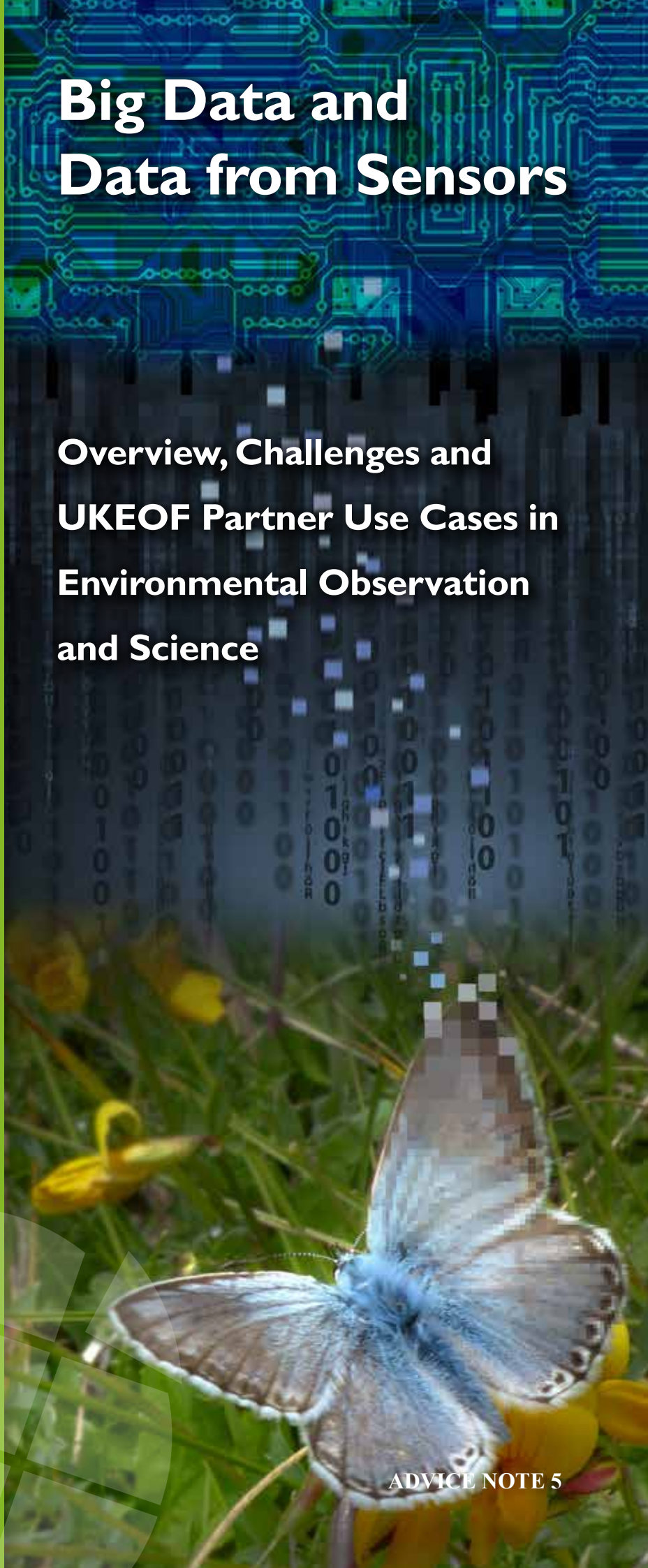
Big Data and Data from Sensors

Overview, Challenges and UKEOF Partner Use Cases in Environmental Observation and Science

Overview

This paper aims to provide a broad range of the UKEOF community and partners with a brief and introductory overview of the current state-of-play of big data, including UKEOF partner use cases in environmental observation and science, and highlights some key challenges in this rapidly emerging and developing space.

Classified as being one of the 'eight great technologies', big data is a rapidly evolving set of concepts and approaches which includes data discovery, collection, (re)combination, mining, analytics and preservation and may be applied to very large, dynamic and complex datasets. The UKEOF Data Advisory Group highlighted that a number of partners are working in this area and that there would be a benefit to bringing their knowledge together. Though aimed at the members of the UKEOF partnership, the paper illustrates in the context of both the providers, users of data and services through relevant use cases, particular technology issues and challenges, which are applicable to the wider environmental science community.



ADVICE NOTE 5

Introducing big data concepts and the importance of software infrastructure for providers and users

Pixabay.com

Digital data on our environment is growing and is made available increasingly rapidly. The variety of data that is available for analysis has also increased and is available in many formats, including real time data from sensors, citizen science and social networks, as well as traditional formats associated with scientific facilities.

There is understandable confusion about the wide range of concepts, technologies and approaches in the burgeoning big data landscape in which all of us now work. The data.gov.uk briefing paper: "Emerging Technologies: Big Data"[1], a horizon scanning research paper by the Emerging Technologies Big Data Community of Interest (December 2014), refers to:

- large volumes of data with high level of complexity
- the analysis used for the data that requires more advanced techniques and technologies to gain meaningful information and insights in real time.

The particular challenges lie in the unstructured nature of much of the environmental data we might want to analyse and rightful concerns about privacy and security in an open data landscape (see Open Data Institute's "Open Data" briefing[2]). In October 2013, the UK Government published its "Seizing the data opportunity: a Strategy for UK Data Capability"[3], which focussed on the three key areas required to increase capability:

- Skills
- Infrastructure, software and collaborative R&D
- Sharing and linking data securely and appropriately.

In order to tackle this landscape, we have seen key trends emerging in the development of cloud computing, new software tools and database systems for large, unstructured datasets and interoperability enabled by use of linked data, standard metadata schema and semantic web based approaches. Some examples of these are presented in the next section. An emerging difficulty is the effective indexing of datasets for data mining and the subsequent representation, visualisation and interpretation of multi-disciplinary datasets, including by non-experts.

Once established, the data feeds and live web services made available can be consumed by Application Programming Interfaces (APIs). These make it possible to integrate multiple data streams in a bidirectional manner, so that updating a record in one database can be seen almost immediately in any other linked database. It becomes possible to listen in to live data feeds and feed decision based algorithms which might lead to an important early warning or alert as well as ensuring records are as up to date as possible, wherever they might be consumed. Scientists and technologists have also been refining analytical tools so that they can process the vast quantities of data in near-real time from sensors. The rise of the 'Smart Cities' concept and 'Digital Government' brings these issues to a head.

There are real concerns around the privacy of data and intellectual property. It is well understood by statisticians that combining multiple anonymised datasets can potentially lead to the unintentional re-identification of individuals. The typical approach to date has proved to be the development of rigorous information security and information governance frameworks, particularly in healthcare, which aim to ensure that only appropriate datasets are available and can be represented from any one service. It is acknowledged that there are still significant challenges to be overcome in order to address this as demand increases for data rich services potentially relating to individuals making use of wearables and other mobile device capabilities. In health, citizens are increasingly expecting to be able to access, e.g. their physical activity data, and relate it to personal medical data where applicable.

The sustainability of these new technologies requires an increasing focus on the provision of support services and monetisation of big data sources. There is a well-known shortage of appropriate expertise, especially within large organisations, to enable this which has led to the launch of new initiatives to accredit the role of e.g. "research software engineers"[4] and seek sustainable funding for them within organisations such as universities. This offers hope albeit demand is already very high for those now in post. Likewise it is typically too costly to build your own in-house infrastructure on an individual project or at an organisation wide level but there is good practice out there in the form of some of the existing data intensive facility and services in place (see use cases below) and under development. The challenge is to make them reusable to a broad range of use cases. In particular we highlight the crucial but underappreciated importance of sustaining the underlying software typically deployed in order to both provision, integrate and analyse datasets. This applies to individuals, projects and organisations and disciplines. Attempts to tackle this include deploying tools for the containerisation of the software, operating system and dependent libraries, suitable deposition in, for example, versioned Github repositories, as well as enabling persistent citation of the software itself, building on the approach used by, for example, DataCite metadata for datasets[5].

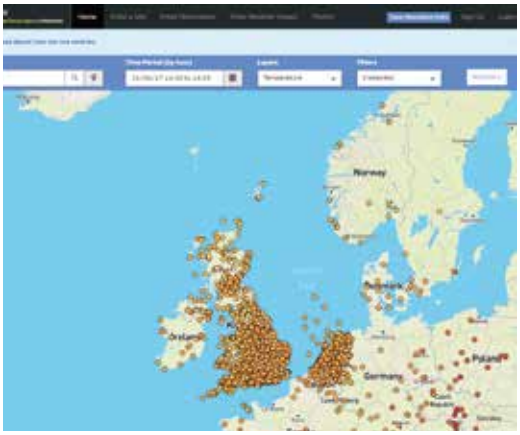
The **key challenges** that emerge and which UKEOF and partners can address, perhaps through dedicated future briefing papers, are:

- Skills including understanding uncertainty (both in the input parameters and model outputs), handling big data (without needing petaflops of computing power), software engineering enabling reuse and making use of big data for meaningful applications
- Infrastructure (including network capacities), the enabling software itself and collaboration
- Sharing, preserving and linking data securely and appropriately, including
  - use of relevant metadata and standard frameworks
  - interoperability as distinct from standards compliance, linked data and the semantic web
  - effective indexing for data mining
- Interpretation of multi-disciplinary datasets and by non-experts
- Benefits to be gained from the interpretation of such large datasets
- Sustainability.

Here we describe some of the data infrastructures being developed in order to tackle these challenges in a UKEOF context. We hope that other such initiatives can benefit from understanding the approaches used and potentially reuse and adapt to their own needs.

WOW Weather Observation Website

The Weather Observation Website (WOW)[6] was launched in 2011 as a repository for citizen weather data. WOW is a website developed on Google App engine and underpinned by a Google Big Table database in the Google Cloud. Users of WOW can submit weather observations from any web enabled device. Observations are stored on WOW and visualised alongside other observations from Met Office official stations and observations from other WOW users, giving a comprehensive picture of the current situation. A user has a unique site ID and PIN to ensure only they can submit data to their site. Data is visible to everyone but can only be downloaded if the site owner grants permission by selecting a tick box in the site details. Photographs can also be submitted alongside observations. Users can even report snowfall through Twitter, this feature will be enhanced upon in the WOW Engine.



Met Office

WOW has not had any marketing, but through word of mouth has expanded to 219 countries as recognised by Google and there are over 8000 registered users who have submitted almost a billion observations since launch. WOW regularly receives close to 30million observations a month and this figure is increasing each month.

Due to its popularity, WOW was approached by other National Meteorological Services to open up the platform for collaborations. In 2012 a pilot was launched with the Australian Bureau of Meteorology that saw the Met Office grant them access to WOW through an Australian sub domain. This has been a huge success and there are over 1300 sites operating in Australia. The Australian example, whilst successful, was also restrictive, as any changes to their branded version of WOW had to be replicated on the Met Office version. In 2014 a collaboration was started with the New Zealand Met Service and the Dutch Met service (KNMI). Using lessons learnt from the Australian collaboration a new API was created to allow for each site owner to make changes without impacting other users.

Building on the success of the above collaborations it was decided to make the whole WOW platform more flexible in terms of the type of measurements recorded, the security of the site and linking to other sites. The latest phase of WOW is called the WOW Engine (launched in June 2016), employing a new development team to radically change the way observations are stored and visualised. It will make use of multiple APIs to enable the collection of a wider range of parameters and data will not be restricted to data that originates from an Automatic Weather Station (AWS). It is proposed to collect data from moving sensor sites such as cars.

As the database will be flexible, it can expand to record new types of variable. Using extensible APIs will allow the WOW Engine to interface with other systems, ingesting data in bulk or exporting a feed of the data to external consumers. As part of the WOW Engine development, the website will be split from the back end database and will continue as a visualisation website that users can interact with. But not everything that is collected in the WOW Engine will have to be visualised on WOW.

It is planned to demonstrate the extensibility of the APIs by building an INSPIRE compliant interface to the WOW Engine. The WOW Engine will aim to be Infrastructure for Spatial Information in the European Community (INSPIRE) and Open Geospatial Consortium (OGC) compliant.

A key lesson from the development of WOW is to make platform storage and collection/dissemination processes as flexible as possible to future proof the infrastructure.

JASMIN Environmental Data Analysis Facility

JASMIN is a very large volume (peta-scale) data analysis facility established to serve the needs of the NERC community as a means to address the challenges of big data. JASMIN provides a centralised resource for hosting key environmental datasets co-located with computational infrastructure to enable processing and analysis.

As data volumes increase the traditional model of users downloading data from a centralised resource to their own client environments is becoming less and less tenable. JASMIN applies the principle of bringing users to the data to address this problem: users are provided with virtual access which directs to the data and computing resources available on JASMIN. For example in the earth observation domain, users have been able to perform full mission reprocessing of data products a task which was hitherto impracticable on institutional based resources alone.

Providing single point access to such a range of datasets fosters interdisciplinary research and collaborative working, essential to modern research and reuse. Along with the provision of shared access to data and processing resources JASMIN provides a community cloud. This enables individual groups to host their own custom applications and services within a tenancy. The ESA-funded OPTIRAD project is one such example. OPTIRAD provides a collaborative environment for land-surface data assimilation enabling researchers to share data and processing and analysis code using the popular IPython Notebook (Jupyter Notebook). Users access an interactive Python command line shell via a web-based interface. This also supports the addition of annotations to code and the incorporation of plots and images in shared notebooks.

The NERC Environmental Research Workbench (ERW) sits as an application layer on the unmanaged JASMIN infrastructure at STFC. Recently released as a prototype system the ERW enables NERC scientists to easily exploit the JASMIN infrastructure and overcome a major hurdle for the many NERC scientists who lack the technical skills and knowledge to exploit high performance and cloud computing infrastructures. It provides an intuitive, user-friendly web interface with a range of tools and services that will help scientists to analyse large and heterogeneous environmental datasets and accelerate science discovery. The Workbench is designed to be extensible, capable of supporting a growing ecosystem of services and increasing community of users for many years beyond its initial development.

The web interface based on an open source content management system and provides a number of pre-loaded tools and services which include:

- Interactive tools and user interfaces running e.g. R-based statistical analysis scripts
- Analytic programming toolkits based on e.g. JupyterHub (IPython Notebook) as a service.

The ERW provides the functionality to enable users to orchestrate simultaneous or multiple runs of these tools and models, the output of which can be shared with individual users or groups. Since the interface is built on an open source content management system sharing of content i.e. tools, data, and output is provided out of the box promoting collaboration and reuse.

The Workbench is designed to be extensible, capable of supporting a growing ecosystem of services and increasing community of users for many years beyond its initial development. Within the ERW environment users can:

- Build new tools and services and make them available
- Configure and customise tools as required (packages, resources)
- Parameterise tools to enable versatility at run time
- Modify existing tools
- Organise tools into categories.

STFC / Stephen Kill

Energy Data in the UK Data Archive


Pixabay.com

Founded in 1967, with the ESRC-funded UK Data Service (UKDS) added in 2012, the UK Data Archive is the UK's largest repository of Economic and Social data. The UKDS is dedicated to providing Users with seamless and flexible access to a wide range of data resources to facilitate high quality Economic and Social Research and Education. Collections curated span well-known national datasets such as the Census and the National Household Survey, international aggregate-level data and smaller studies with themes including ageing, housing, labour, environment and energy.

The UK Data Service is extending and expanding its capacities to facilitate the curation and analysis of new and novel forms of data, and is engaged in a major project to develop big data architecture for social science. Challenges to archival processes include not only the introduction of new technology, necessitated by the volume of data in datasets, but also changes in cognition, policies and workflows. While the UKDS was working with digital objects, at times necessitating complex analyses, novel requirements included linking datasets and work with social media. The aim has been to develop discipline-agnostic, generic systems and tools.

UKDS has focussed on a big data test case and proof-of-concept utilising household energy consumption data collected from smart meters throughout the UK. Electricity and gas measurements are taken on a continuous 30-minute basis, recordings which are provided alongside geographic information. Although it should not be possible to identify any individual from the data, the data is classified as safeguarded. There are several strands to this project. Firstly, research into the data itself implementing a data-driven rather than a hypothesis-driven model through exploratory data analysis and visualisation techniques. Secondly, the development of appropriate IT infrastructure creating, insofar as is possible, a big data service for the social science community. Thirdly, scoping interest and requirements within the social science community, demonstrating potential for big data using our research as a proof of concept model.

In developing these systems and research, helping to shape recommendations at national and international levels, the Open Data Platform (ODP) data lake approach has been adopted. We are utilising the Hortonworks Data Platform, with its associated data storage, processing and analytics components to showcase the power of the ODP model to provide a cost-effective and scalable framework. Students and researchers working on this platform are exposed to industry-standard data tools. Incorporating our 'Five Safes' framework, we are developing hybrid architecture, with cloud and on-premises installations, to serve both secure and non-secure data services.

References

1. https://www.gov.uk/government/publications/emerging-technologies-big-data
2. http://theodi.org/what-is-open-data
3. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/254136/bis-13-1250-strategy-for-uk-data-capability-v4.pdf
4. http://www.rse.ac.uk
5. Jones C., Matthews B, Gent I, Griffin T, Tedds J. (2016). *Persistent Identification and Citation of Software,* International Journal of Digital Curation, Vol. 11, Iss. 2, 104–114 http://dx.doi.org/10.2218/ijdc.v11i2.422
6. http://wow.metoffice.gov.uk

ACKNOWLEDGEMENTS

Thanks to Liesbeth Renders, Bronwen Williams and Vicky Morgan for UKEOF oversight and facilitation. Images courtesy of Pixabay.com unless otherwise stated.

CONTRIBUTING AUTHORS

Jonathan Tedds (Editor, University of Leicester) – jat26@le.ac.uk
Mike Brown (CEH) - mjbr@ceh.ac.uk
Phil Kershaw (STFC) - philip.kershaw@stfc.ac.uk
Dom Lethem (Met Office) - dom.lethem@metoffice.gov.uk
Ben Wright (Essex) - bwrighp@essex.ac.uk

UKEOF
ENVIRONMENTAL OBSERVATION FRAMEWORK

UKEOF works to improve coordination of the observational evidence needed to understand and manage the changing natural environment. It is a partnership of public sector organisations with an interest in using and providing evidence from environmental observations. Contact us at office@ukeof.org.uk

UKEOF PARTNERS

Department for the Environment , Food and Rural Affairs, Department of Energy and Climate Change, Economic and Social Research Council, Environment Agency, Forestry Commission, Joint Nature Conservation Committee, Met Office, Natural England, Natural Environment Research Council, Natural Resources Wales, Northern Ireland Environment Agency, Scottish Environment Protection Agency, Scottish Natural Heritage, The Scottish Government, UK Space Agency, Welsh Research Environment Hub, Welsh Government.

ACRONYMS

- API – Application Programming Interface
- ERW - NERC Environmental Research Workbench
- ESA - European Space Agency
- INSPIRE - Infrastructure For Spatial Data in Europe
- JASMIN - data analysis infrastructure and services for the NERC community hosted by STFC
- OGC - Open Geospatial Consortium
- OPTIRAD - Optimisation Environment For Joint Retrieval Of Multi-Sensor Radiances (a European Space Agency funded project to host a collaborative research environment on the JASMIN cloud
- STFC - Science & Technologies Research Council
- UKDS - UK Data Service

STFC / Stephen Kill

This is one of a series of advice notes prepared by the UKEOF Data Advisory Group.

Front cover image: CEH / Andrew Sier. Graphic sources: Pixabay.com & A. Sier

Published 2017

www.ukeof.org.uk

@UKEnvObs