How do we manage the data deluge?

The “principles” of good data management

Dr David Morris

Cefas
Presentation "metadata"

information about the presentation

- **Start** – deluge
- **Next** – Cefas
- **Then** – UK-EOF on Data Management
  - With stuff from me on metadata
- **Followed by** - UK-EOF on Citizen Science
  - With stuff from me on metadata
- **Finally** – having shown you need to care, here’s where you can share
Data deluge
What happens when the world is plugged into the network?

Data overload
Back in 2010 Google chief executive Eric Schmidt noted that the amount of data collected since the dawn of humanity until 2003 was the equivalent to the volume we now produce every two days.

The 2004 size of the Internet was estimated at 5 trillion terabytes, or 5 exabytes.

In 2013 internet data, mostly user-contributed, will account for 1,000 exabytes. An exabyte is a unit of information equal to one quintillion bytes.

- Each engine of a jet on a flight from London to New York generates 10TB of data every 30 minutes
- In 2013 internet data, mostly user-contributed, will account for 1,000 exabytes. An exabyte is a unit of information equal to one quintillion bytes
- Open weather data collected by the National Oceanic and Atmospheric Association has an annual estimated value of $10bn
- Every day we create 2.5 quintillion bytes of data
- 90% of the data in the world today has been created in the past two years
- Every minute 100,000 tweets are sent globally
- Google receives two million search requests every minute

http://www.bbc.co.uk/news/technology-23253949
Après le deluge, moi!

the BIG BLUE wet thing
• From gene sequences, through ear bones (50,000 a year) and pictures of slices of sea bed, and diseases, and blooms, and radiation and, and, and …..

• Using everything from electronic tags, Smartbouys, Waveriders, autonomous gliders to market fish samples and questionnaires (now on iPads) and, and, and ….

• Crown Estate (a £9Bn entity) holds some 50 TBytes of data, Cefas (a £50m entity) 40 TBytes and growing [excludes one current project Marine Protected Areas, 60 TBytes and growing] and its all (pretty much) numbers, not the disorganised text, images and “stuff” of most “Big Data”

• We are a Crown Body so at the coal face (fracking pad?) when it comes to making publicly funded data accessible and available (a separate talk entirely) and we were doing Citizen Science before it was called that (both with thermometer wielding citizens and fishermen)
Coastal Temperature Network - 1890, 1900, 1920, most 1970 on (volunteer observers)
Can citizen science be used to deliver policy-relevant marine data?

Kieran Hyder¹, Victoria Bendall¹, Stuart Hetherington¹, Julia Hunt¹, David Pearce¹, John Pinnegar¹, Dave Sivyer¹, Bryony Townhill¹ & Simon Keeble²
¹ Centre for Environment, Fisheries & Aquaculture Science, Lowestoft, UK & ² Blue Lobster IT Ltd, Norwich, UK

The Policy Challenge

Large datasets are required to support increasingly complex marine policy (e.g. Marine Strategy Framework Directive). However, data collection is expensive and funding is limited.

It is important to look for novel ways of obtaining and processing data.

Can Citizen Science Help?

Citizen science has great potential to add to the marine evidence base. To assess the potential we will:

1. Understand current activities and engage with existing initiatives.
2. Test two potential case-studies involving divers and anglers.
3. Make recommendations about future use of citizen science.

Temperature Profiles from Scuba Divers

UK divers make about 1.6M dives every year recording temperature profiles on their computers.

Many dives are shared on diver social networking sites.

Compile temperature data

Benefit for science and divers

Impact of climate change on young fish
Validating hydrodynamic models
Information for divers

Anglers Tagging Sharks and Rays

Many sharks and rays are endangered.

Electronic tags provide vital information to help conserve stocks, but tagging programmes are expensive.

Anglers can help

Anglers are already involved in tagging fish. We will work with a small number of anglers to deploy electronic tags.

Benefit for science and anglers

Engagement with anglers and learning from each other
Better scientific understanding and conservation of sharks and rays
More and bigger fish to catch

This work is funded by the Department for Environment, Food and Rural Affairs under the contract MF1230 “Citizen Science Investigations - empowering the public through the use of novel technologies to collect policy-relevant marine data”

kieran.hyder@cefas.co.uk
Principle 6

Public Information is published

Public Information includes the objective, factual, non-personal information on which public services run and are assessed, and on which policy decisions are based, or which is collected or generated in the course of public service delivery. Public information should be published, unless there are overriding reasons not to.

Crucially, this principle goes beyond the minimum requirements imposed by legislation. It advocates a proactive approach to publication of information – i.e. presenting, formatting and promoting information in useful formats for wider consumption, without it needing to be specifically requested or mandated in legislation.

Note that publishing information to the public also requires consideration of the practical channels by which this will actually be achieved. This includes the establishment of internal publication processes, the use of publication hubs (e.g. data.gov.uk), as well as potentially relationships with ‘third party’ information intermediaries.

Clearly the desire to publish information does need to be balanced against constraints which may prevent this. Exclusions would include, for example, personal information, information which can compromise privacy, commercially and legally privileged information, and information that is required to maintain security.

Principle 7

Citizens and businesses can access information about themselves

Citizens and Businesses should be able to access information about themselves, along with an explanation of how it is used. This may be either on request or, preferably, by making it available by default. In effect, such information should be considered as belonging to the citizen, although entrusted to the care of a public body.

Note that this principle goes beyond the minimum requirements imposed by legislation. It advocates a proactive approach to allowing citizens to access information about themselves, without it necessarily needing to be specifically requested or mandated in legislation. This might be achieved, for example, by making it securely available online. Consideration needs to be given to both viewing and, where appropriate, to performing transactions such as updates (for example to correct inaccuracies).

Clearly the desire to make information available does need to be balanced against constraints which may prevent this. Exclusions would include, for example, legally privileged information, and information that is required to maintain security.
The principles of good data and information management

6. Public information is published
5. Information is reused
4. Information is standardised and linkable
3. Information is fit for purpose
2. Information is managed
1. Information is an asset

The Chief Information Officers Council has identified seven principles which build naturally into a hierarchy, as depicted in the diagram above. For example, it is unlikely that information can be reused (Principle 5) unless it is also valued, managed, fit for purpose and standardised (Principles 1-4).

**Information is valued asset**
Information should be understood and valued as much as other organisational assets such as buildings, machinery, people or money.

This principle is the foundation for what follows and highlights the need for information to be valued in the same way as these other types of asset. It is important to note that the true value of information lies not just in its original purpose but in its potential to be reused for other purposes.

**Information is managed**
Information should be managed – stored, protected and exploited – according to its value.

Data and information managers need to consider the whole lifecycle of the information, from identification of need, creation, quality assurance, maintenance, reuse and ultimately to archiving or destruction once the information has ceased to be useful.

A range of best practices need to be in place, for example to ensure appropriate availability and integrity, avoid loss and ensure continuity across technology upgrades. It is particularly important that personal data are adequately protected. Information also needs to be governed as it moves through its lifecycle, for example to make sure it’s always clear who is responsible for it (be an identifiable owner), and to comply with relevant legislation and regulation.

The consistent assessment and ownership of these information risks is another important consideration when managing data and information.

The organisational culture must support best practice in data and information management, and make sure everyone responsible for processing these business assets is professionally qualified and appropriately skilled. This principle therefore also includes the processes, roles, responsibilities, training, and organisational structure and culture needed to ensure the effective and efficient use of information.

**Information is fit for purpose**
Information must be good quality and fit for both its primary purpose and potential secondary uses. It will not always be possible for the originator to foresee secondary uses, so it is important that the quality of the information is communicated consistently so future users can decide if it is suitable.

Quality includes factors such as accuracy, validity, reliability, timeliness, relevance and completeness. The quality of data and information should also be regularly monitored to ensure that they at least meet the levels that have been assessed as necessary for their purposes.

A further aspect of this principle is to consider aligning the supporting technical platforms and format with how information will be used. For example, if information is likely to be needed for online statistical analysis, it won’t be appropriate to store it in a system or format that is only accessible to the originator, on backup tapes or unstructured PDF format.

This principle depends on information to be perfect, only that it is the right quality for its intended use and that its quality characteristics are clear to future users.

**Information is standardised and linkable**
There will be many more opportunities for exploiting information if it is available in standardised and linkable forms.

Standardisation is important for structured information such as dataset definitions, and unstructured information such as metadata tags applied to documents. Standardisation within an organisation is important for staff to fully exploit the information; if an organisation uses widely accepted open standards it will unlock even more value for other users.

Even further value can be unlocked if information can be linked. A good example is document references and citations that allow the reader to draw on a wealth of associated information (this is the basis of the semantic web). For example, tagging spending information with an authoritative code for the organisation would allow it to unambiguously linked with details of the organisation itself and third-party information about that organisation (eg service satisfaction measures).

**Information is reused**
Information is even more valuable if it can be used more than once or for more than one purpose. A good data manager will proactively look for opportunities for reuse.

These could include:
- Internal reuse – making the most of information for its primary purpose and identifying secondary uses. For example, operational data can sometimes be reused to support performance improvement or research.
- External reuse – sharing information with other organisations, either within the public sector or with private businesses and citizens.
- Holding master data – ensuring an organisation’s data is the only authoritative source for business (information i.e. an authoritative list of organisation codes), which is nominated, maintained and promoted as such.

Reuse involves considering what information an organisation can make available to others, and looking at how an organisation might reuse information held by others.

Within this principle strongly encourages reuse, it is important to appreciate that reuse does require a careful risk-based judgement to be made with regard to exploiting versus protecting information, as well as consideration to the costs and benefits involved, and any rights or other commercial considerations.

Information which initially appears unsecure may be reusable if it can be reformatted. For example, operational information that identifies individuals can be anonymised or aggregated and then be of wider value. Also, in cases where the partner organisation is known beforehand, concerns over security or privacy can sometimes be mitigated by means of negotiation, joint working and data-sharing agreements.
The principles of good data management

1) Information is a Valued Asset

2) Information is Managed

3) Information is Fit for Purpose

4) Information is Standardised and Linkable

5) Information is Re-used

6) Public Information is Published

7) Citizens and Businesses can Access Information About Themselves

As used in the latest Draft of the Defra Network Knowledge Strategy Embedded Public Sector Information Principles

Separate talks in themselves
Scientists leading or participating in citizen science projects are primarily interested in the scientific outputs. They may be professional scientists or leaders or coordinators of natural history groups, environmental charities, governmental agencies or non-governmental organisations (NGOs).

A participant is an unpaid person who takes part in a project by helping to define its focus, gather or analyse data – a ‘citizen scientist’.

http://www.ukeof.org.uk/co_citizen.aspx
### Before you start
- Is citizen science the best approach? - 2
- Choose a citizen science approach - 4
- Citizen science flowchart - 6

### First steps
- Establish project team - 7
- Define project aims - 8
- Identify funding and resources - 9
- Identify and understand target participants - 10

### Development phase
- Design the survey or scheme - 12
- Consider data requirements - 14
- Consider technological requirements - 16
- Develop supporting materials - 17
- Test and modify protocols - 19

### Live phase
- Promote and publicise the project - 21
- Accept data and provide rapid feedback - 22

### Analysis and reporting phase
- Plan and complete data analysis and interpretation - 23
- Report results - 24
- Share data and take action in response to data - 25
- Evaluate to maximise lessons learned - 26

Separate talks in themselves At least!
Citizen science works best when:

- the project is an efficient and enjoyable way to gather and analyse the required dataset;
- the quality of the scientific data generated is measurable.
Key EOF Data messages (1)

• Just because you (and your ‘citizens’) can measure something (collecting a lot of data, everywhere, all the time) doesn’t mean you should.

• If you must, **think** hard, **plan** ahead
  – what will it be used for
  – how will it fit with other ‘professional’ data users and resources
  – How will you answer quality questions (at source and in use)
Key EOF Data messages (2)

• Just because the source is ‘free’ and ‘amateur’ doesn’t mean you don’t have to curate the data properly, including security, access, use and licences (a potential world of pain that can be mitigated by careful planning and good systems so the machines do the work, not you)
Key EOF Citizen Science Messages

Consider data requirements, storage & analysis

Accept data and provide rapid feedback

Share data and take action in response to data
Key Personal Messages (1)

• from my experience in putting a really whizzy system into Cefas (>1/2 way there)

Wave data milestone recorded: 1 million hours and counting

Reference: 06-13
05 August 2013
Key Personal Messages (2)

• Pretty much (and they are all linked)
  – PEOPLE are the “problem”
  – Doing things to data/metadata retrospectively is BAD (and expensive)
  – Systems in the SERVICE of people is the “answer”
  – Engaged and enthusiastic and informed people are GOOD
  – Data Management is DULL and a means to an end
    • So let the machines do the work
  – The word METADATA is a universal SOPORIFIC however, it is vital
Not everything that counts can be counted, and not everything that can be counted counts.

Or was it someone else? If only the quote had metadata!
“metadata” helps you get things right


- *It would be nice if all of the data which sociologists require could be enumerated because then we could run them through IBM machines and draw charts as the economists do. However, not everything that can be counted counts, and not everything that counts can be counted*
Whatever data you collect you will need metadata

- **Data without metadata is worse than useless;** it's expensive, you have to manage it, you have to look after it but you can't really use it properly and confidently, and you may have to delete it, which can be complicated.

- **So the trick is** to get metadata generated automatically where you can; positions are easy, so is date/time, data quality is more difficult as that requires your user to do something – focus on making that simple and easy.
Metadata:
Title, Description, Lineage

- **Title** – YOU supply this as part of the project
- **Description** - YOU supply this as part of the project
- **Lineage** – YOU supply MOST of this as part of the project, BUT there are bits you can’t really supply but the ‘citizen scientist’ can
Lineage = Quality

• Back to design – can the app ask the user questions that will assist in determining the quality?

• Can photographs help the user (and you) determine something, eg the species?
  – You supply images (initial work up front)
  – They take images (lots of work for you ‘later’)

• What does the user ‘think’ – ask the right questions and presses of the button at the time writes the lineage statement for you
Key Citizen Science Messages

- Consider data requirements, storage & analysis ✔
- Accept data and provide rapid feedback ✔
- Share data and take action in response to data ?

---

UKEOF
ENVIRONMENTAL OBSERVATION FRAMEWORK

SEPA
Scottish Environment Protection Agency

Cameras
A CO-ORDINATED AGENDA FOR MARINE, ENVIRONMENT & RURAL AFFAIRS SCIENCE

Living With Environmental Change
• If you have
  – The data
  – The metadata
  – Assured quality (NOTE this can be low, just say so!)
  – Clear licence conditions (easy these days – see Open Data initiatives)
  – Then there are all sorts of places for data, Citizen Places, national places, International places
Documenting, archiving, and sharing your data has never been easier. Use DataUp to create metadata, perform a best practices check, and post your tabular data to a repository for others to use.

Sign in to get started

Use DataUp for these tasks:

1. **upload**
   - Upload tabular data to DataUp

2. **best practices**
   - Perform a check for data best practices

3. **describe data**
   - Describe your data with standardized metadata

4. **generate citation**
   - Generate a citation with a persistent identifier for your data

5. **post to repository**
   - Post your file to a data repository

[https://redirect.dataup.org/](https://redirect.dataup.org/)
and especially for the BIG BLUE wet thing

British Geological Survey (BGS) for seabed and sub-seabed geology,geophysics data more>>

The Archive for Marine Species and Habitats Data (DASSH) for flora, fauna and habitat data more>>

British Oceanographic Data Centre (BODC) for water column oceanographic data more>>

United Kingdom Hydrographic Office (UKHO) for bathymetry data more>>

The Met Office for marine meteorological (metrocean) data more>>

Centre for Environment, Fisheries & Aquaculture Science (CEFAS) and Marine Scotland Science (MSS) for marine fisheries data more>>

Archaeology Data Service (ADS) for marine historic environment data more>>

If you are unsure which data archive centre to submit data to please contact MEDIN Enquiries.
Data + Metadata = Good Data

• Plan to get the machines to do most of the work – from start to finish – from collection to sharing to analysis

• Data without metadata isn’t really worth the effort and is certainly less valuable/useful
  – Metadata helps you turn data into a managed, valued, fit for purpose, reusable asset (4 out of 7)
  – Metadata in all 5 Citizen science phases, planned, from the start, as you develop, during the live phase (capture) and in analysis

Metadata is an unsung hero of the modern world, the plumbing that makes the information age possible.

Metadata is a tool that enables the information age functions performed by humans as well as those performed by computers.