Open Data

A presentation from the "Technical & resources" section of the NBN Data Standards & Tools Steering Group (2006*) for the BRC 2024 National Recording Schemes Meeting *DF Bulletin #72

"There are known knowns, things we know that we know; and there are known unknowns, things that we know we don't know. But there are also unknown unknowns, things we do not know we don't know."[Rumsfeld]

Countries are thirsty for summarized data and insights for policy-making but we are running short of tools (Martinez, 2023)

"UK has a quarter of all Diptera records on GBIF" Chris Thompson (Smithsonian), 2009

Known knowns

Widely known about in our sector of science of course due to the work by NBN, BRC and newsletters by NFBR, GIGL & TVERC etc.. Not so much outside it; New Scientist never but British Wildlife are crying out for new writers if anyone's game.

Open Data in other sectors is a popular subject, read Ritchie's "Science Fictions" and mentally add a chapter from our sector.

Known unknowns

Online Open Data publishing is an aspiration, it's far from comprehensive though. The gaps are widespread and huge, they need comprehensively reviewing so as to clarify our objectives:

Recording Schemes

Half of our Diptera schemes publish Open Data, amounting to ~15% of total available digitised species occurrences.

- Buglife analysed 45M invertebrate records from Recording Schemes, way more than the combined silos of BRC & NBN (which overlap.)
- What arrangements to support archive and security have been made for all the unFAIR data?

Journals

Not many journals or reports in our sector embrace the Open Data principle by specifying an OD source of occurrences in articles. Only PlosOne includes it in their guidelines for authors, Zootaxa doesn't demand it. Smaller journals may sometimes attempt FAIR principles but they are very rarely encountered in ento journals. We're decades behind the medical sector. Guidance can be found at http://tinyurl.com/5f9bz2m8

- Since we advocate the scientific value of OD and do most of its gathering then shouldn't we be pressuring all journals to mandate its use in published articles?

Local Recording

What's the proportion of records that are retained locally and not submitted to NBN Atlas? There used to be a financial incentive from English Nature for LRCs to upload. Charles Roper detailed data flows and I reported to ALERC on RS some years ago,

- What's the data flow picture now?
- What's the current status of our biological recording software?

Collections

Lifetime works by naturalists end up in museums (or skips in at least one famous case.) Some of this may have achieved Open Data status during the collector's lifetimes but a good proportion has not. Collection datasets can be found on NBN Atlas, Derek Lott's coleoptera is an example, but all collections should be uploaded. This process began in earnest in 2023 with the DISSCO project (see comment by Open Data Institute at https://tinyurl.com/4hb3872t)

- A summary of uploaded Open Data datasets across the entire museum sector and an indication of progress
- A review/catalogue of collections awaiting or undergoing processing

The Naturalist's Toolkit

The following processes are pretty much common to all naturalists, the latter half of them in particular by those organising Recording Schemes.

	Collection	Traditional + iRecord & iNaturalist. At least 8 different methods
	Collation	Desktop applications: QC issues & verification
	Management	Desktop applications such as iMatch, Recorder versions, MapMate
	Analysis	various such as Biogeography, Phenology. use of R.
	Dissemination	websites with ID keys, guides etc.
	Publication	includes both journals & open data publishing + DwC issues

Most of the tools we use find their way into that list somewhere so it could be used as the basis of a RS survey. Online systems have achieved inroads into some of them but by no means all. For example not your image collection or biogeographical, trend and other analyses.

- Huge rise in website costs to Recording Schemes (£3,000+ each) following the Natural History Museum's freezing of an enterprise scale (~£3M) service (Scratchpads http://tinyurl.com/2u95hstb) No marketplace equivalents or alternatives available. Loss of valuable online storage space for guides and no funding.
- We've also lost the FSC Biodiversity Forum which supported their online Identikit key system (which still works)
- Incoherent taxonomic support once you stray out of the UK
- Darwin Core utility allowing conversion from spreadsheets available to professionals or R6 users, not to most RS organisers though. (see Mesibov on GBIF, e.g. at http://tinyurl.com/bdzdaeja)
- No dedicated or suitable Electronic Document Management Systems or Citation Managers

Expeditions

From brief rambles by Natural History Societies, through surveys conducted by LRCs and Recording Scheme groups to international expeditions organised by UK museums.

- Can we locate the museum expedition Open Data datasets?

Dipterists Forum have published numerous Open Data datasets from annual expeditions (http://tinyurl.com/5n7wdx4k) since 1998, with BRC help the latest few are down to a fine art.

FAIR Badges DwC-A FAIR ☑ = accreditation?

Love badges? There's a system for obtaining those, detailed by Jorrit Poelen at http://tinyurl.com/aw4znkeu Test it on your datasets to get them for your website etc. (sample mine.) Perhaps the NBN partner pages could display those for our datasets.

Unknown unknowns

Engage the recording interests of naturalists via the Recording Schemes and we'll find out (iRecord, iNaturalist, desktop systems etc.)

Making Data Work Harder

Open Data advocacy has been the subject of numerous articles in various forums over the past 25 years. The above summarises several of the most recent issues relating to Recording Schemes and Field Expeditions.

Darwyn Sumner: GBIF Biodiversity Open Data Ambassador
[Organiser of three Recording Schemes]

With thanks to Chris Thompsin (Smithsonian), Giselle Sterry (NBN), Sophie Ratcliffe (NBN), Sarah Whild (NFBR), Phil Brighton, Robert Mesibov & Jorrit Poelen