A standardized data layout for archiving

Allan Reese and Grant Stentiford
Centre for Environment, Fisheries and Aquaculture, Weymouth
cefas.co.uk

Why archive? Why standardize?

Data are expensive to collect but have lasting value. Climate change, chronic effects of pollution, and biodiversity change all prompt re-use of data and samples collected years earlier for other purposes. Data if not archived are lost to science, but re-used data may be misinterpreted when methods and assumptions are not described in sufficient detail - it's easy to assume "everyone knows that." Problems arise from acronyms, local jargon, made-up codes, implied dittos, and mixing data with subtotals etc.

We must all manage storage of data and samples more effectively. Digital storage is replacing visible storage of physical notebooks. Constraints imposed by the immediate IT environment may be mistaken for features essential for long-term storage. Files can be lost in cyber-space.

Adopting a standard layout eases all the tasks of depositing data, cataloguing files so they don't get lost, and understanding the content for appropriate re-use.

Standards for Metadata

Metadata is the key to data, but standards from computer science describe storage structures rather than meaning. The "Dublin Core" (NISO Standard Z39.85-2001), for example, lists required fields such as "title" and "description", but leaves their content to the individual. This risks omitting information essential to new users outside the research group.

Social scientists are more used to data archiving. The ESRC requires grant holders to deposit their data with the UKDA (www.data-archive.ac.uk). Archive staff help to ensure that "data are deposited to a standard that enables them to be used by a third party, including the provision of adequate documentation. Consent, confidentiality, ethical and legal issues are considered, included in the project management plan and addressed before data collection starts."

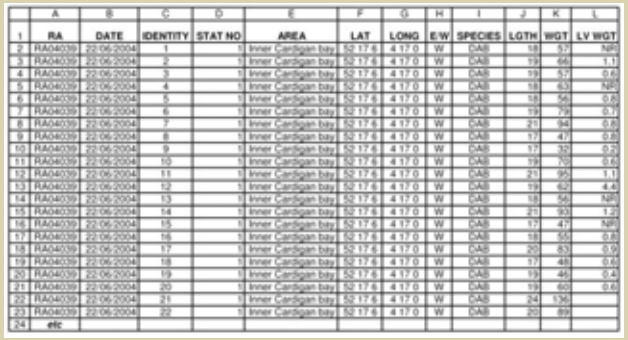
Our own project clarified roles for data ***managers*** and data ***collectors.*** Managers are IT specialists who guarantee the physical security and continued readability of files, but take on trust that the contents make sense. Collectors - scientists or administrators - are perhaps too close to their data to notice the assumptions and feel more pressure to meet their own objectives than think of alternative uses.

A standard layout for data collectors

Like it or not, Microsoft Excel is widely used for data handling. Its help guidelines on LISTs are clear and relevant. Data relating to more than one level may imply an Access database, but the extra complication is often not justified. Scientific data tend to be stable once checked, and transaction processing is not a good model. However, both programs are so flexible that idiosyncratic use is the norm, making data structures hard to recognise.

We propose a ***three-step procedure*** for data submission. ***Raw data*** in columns form a single table (or view from a relational database); the ***codebook*** describes each column; and the ***metadata*** description uses fields from the Dublin Core. These three documents could be submitted as text files (eg CSV ASCII) but we use the familiarity of Excel to provide a template of three sheets in a workbook, as shown. This example from fish-disease monitoring is at the stage where it should be checked by the data manager or a third party.

Notes explaining the system in more detail as a standard operating procedure (SOP) are available from the first author.



	A	B	C	D	E	F	G	H	I	J	K	L
1	RA	DATE	IDENTITY	STAT NO	AREA	LAT	LONG	E/W	SPECIES	LGTH	WGT	LV WGT
2	RA04039	22/06/2004	1	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	18	57	NR
3	RA04039	22/06/2004	2	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	19	66	1.1
4	RA04039	22/06/2004	3	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	19	57	0.6
5	RA04039	22/06/2004	4	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	18	63	NR
6	RA04039	22/06/2004	5	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	18	56	0.6
7	RA04039	22/06/2004	6	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	19	79	0.7
8	RA04039	22/06/2004	7	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	21	94	0.8
9	RA04039	22/06/2004	8	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	17	47	0.5
10	RA04039	22/06/2004	9	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	17	32	0.2
11	RA04039	22/06/2004	10	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	19	70	0.6
12	RA04039	22/06/2004	11	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	21	95	1.1
13	RA04039	22/06/2004	12	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	19	62	4.4
14	RA04039	22/06/2004	13	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	18	56	NR
15	RA04039	22/06/2004	14	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	21	93	1.2
16	RA04039	22/06/2004	15	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	17	47	NR
17	RA04039	22/06/2004	16	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	18	55	0.6
18	RA04039	22/06/2004	17	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	20	83	0.9
19	RA04039	22/06/2004	18	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	17	48	0.5
20	RA04039	22/06/2004	19	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	19	46	0.4
21	RA04039	22/06/2004	20	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	19	60	0.6
22	RA04039	22/06/2004	21	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	24	136	
23	RA04039	22/06/2004	22	1	Inner Cardigan bay	52 17.6	4 17.0	W	DAB	20	89	
24	etc											

*1) **Raw data** sheet. Table starts at top left (cell A1). No random text or other workings round or in-between the data. All cells are values, not formulae. First row used for one-word unique column headings (names), supplied by the user. Data table is rectangular, so repeat values must be copied onto each row and missing values flagged with non-blank value. The Manager should refer the data back to get the blank "LV weights" changed. Excel allows mixing numbers and text cells within a column; most statistics programs will not.*

*2) **Codebook sheet.** Our standard column headings in the template set out what is needed to expand each column heading in the raw data. The name, which may be cryptic, is explained in a* label *and has the* type, storage *and* meaning *of values explained. Details of acceptable values or ranges can be applied to the data collection process, so that values are checked as acceptable and consistent (no pregnant males unless sea-horses) immediately on entry. The* reference *is vital, so that studies that use the same name with different definitions are not naively merged.*

*3) **Metadata** sheet. Several fields have been left for the data manager to fill. The geographical bounding box can be deduced from the data, and the Manager will know how to round that to catalogue standards. One general question is how to handle datasets containing many languages, which might be as simple as a list of people's names.*

Question

Should it be Defra policy to archive data centrally, within each agency, or with an outside contractor such as UKDA?

This work was supported by Defra under contract F1171