

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.



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

© Crown copyright 2007

A standardized data layout for archiving



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

	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	0.8
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	0.8
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	0.8
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	0.6
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.

	A	B	C	D	E	F	G	H
1	Name	Label	Type	Format	Units range	Missing values	Codes	Reference
2	RA	Research Accession number allocated to sample set	String	Combines RA prefix, then sequential number for sample set	Not applicable	No missing values	Not applicable	Sequential numbering system for sample sets. Allocated from the RA database in Pathology and Parasitology team in Weymouth. Samples collected under each RA are stored as wax blocks and slides in Weymouth
3	DATE	Date of sample collection	Date	DDMMYYYY	Cannot be later than today	-1 if day unknown, -1 if month unknown, -1 if year unknown	European date format	Not applicable
4	IDENTITY	Sequential number allocated to individual specimen with RA	Integer	Sequential number (1 to n)	Standard units	No missing values	Not applicable	e.g. Individual fish would be allocated an RA number and a specific sequential number identifier
5	STAT NO	Station number recorded from ship log in specific year	Integer	0-999 (almost never more than 999 stations per cruise)	1-999	No missing values	Not applicable	Sequential task undertaken on cruise. Do not cross-reference to any other cruise. Consecutive stations may be at same or next location.
6	AREA	Free text name of sampling site	Text	Name of site	Not applicable	No missing values	Not applicable	Not applicable
7	LAT	Latitude of specific station number	string	e.g. 53 28 220 refers to 53o28' 220"	Check units locate sampling site to UK marine environment	No missing values	Not applicable	A latitude of 53 28 220 refers to 53o28' 220 N (or, 53 degrees, 28 minutes, 220 seconds North)
8	LONG	Longitude of specific station number	string	e.g. 03 19 200 refers to 03o19' 200"	Check units locate sampling site to UK marine environment	No missing values	Not applicable	A longitude of 03 19 200 refers to 03o19' 200" (or, 03 degrees, 19 minutes, 200 seconds). For easting or westing reference, check E/W variable
9	E/W	Easting or Westing	string	A	Only defined codes	No missing values	E = easting W = westing	Apply easting or westing to coordinate. Correct location of sampling site can be cross referred to station number for year (STAT) and/or name of site (AREA).
10	SPECIES	Species generic name	string	AAA	Not applicable	No missing values	Not applicable	Refer to ICES/Cefas species codes URL?
11	LGTH	Body length	Real value	two digits	Whole centimeter. Expected range 5 to 40cm	May be recorded as NR if mistakenly not recorded	Not applicable	Values above 0.51 of unit rounded up to nearest cm, values below 0.50 of unit rounded down to nearest cm. Length measured from tip of snout to tip of tail fin
12	WGT	Body weight	Real value	two or three digits	Whole gram	May be recorded as NR if mistakenly not recorded	Not applicable	Values above 0.51 of unit rounded up to nearest gram, values below 0.50 of unit rounded down to nearest gram. Weight measured on calibrated weighing balance
13	LV WEIGHT	Liver weight	Real value	one or two digits and one decimal	gram and decimal gram	Recorded as either NR (not recorded due to mistake) or blank cell (not recorded for this fish)	Not applicable	Liver weights generally recorded for first 20 fish at each new site visited within RA. NR refers to a sample mistakenly not recorded for LW while a blank cell refers to sample not required for this fish

	A	B
1	Title	RA04039 Disease data for dab: RV Cefas Endeavour, cruise CEND 07/04
2	Dataset language	English
3	Topic category	(from IMET)
4	Subject	dab, Limanda limanda, fish diseases, liver pathology, National Marine Monitoring Program, NMMP, CSEMP, UK marine, North Sea, Irish Sea, English Channel
5	Dataset reference date	22/06/2004 to 9/7/2004
6	Originator	Grant Stentiford, Pathology and Parasitology, Weymouth laboratory
7	Lineage	Annual marine fish disease monitoring data. Data for other years available
8	West bounding coordinate	Entry by data manager
9	East bounding coordinate	Entry by data manager
10	North bounding coordinate	Entry by data manager
11	South bounding coordinate	Entry by data manager
12	Extent	Irish Sea, English Channel and North Sea
13	Vertical extent information	Not applicable
14	Spatial reference system	Entry by data manager
15	Spatial resolution	Entry by data manager
16	Spatial representation type	Entry by data manager
17	Data format	MS Excel data table
18	Frequency of update	Entry by data manager
19	Access constraints	Entry by data manager
20	Use constraints	Entry by data manager
21	Additional information source	Fish disease summary data and interpretation published in regular Marine Environmental Monitoring Reports published by Cefas
22	Abstract	This data table contains marine field data collected on the RV Cefas Endeavour, cruise CEND 07/04 in 2004. The raw data represents individual dab (Limanda limanda) examined for a range of externally visible diseases (recorded on board ship) and 31 specific liver pathology categories recorded from material collected on board ship and analysed in the laboratory. Liver samples for each specimen have been retained as stained microscope slides (H&E stained) and also within wax blocks. In addition, frozen samples may also have been taken from individuals (please refer to codebook and raw data). All specimens collected and stored under Research Accession (RA) 04039.

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.