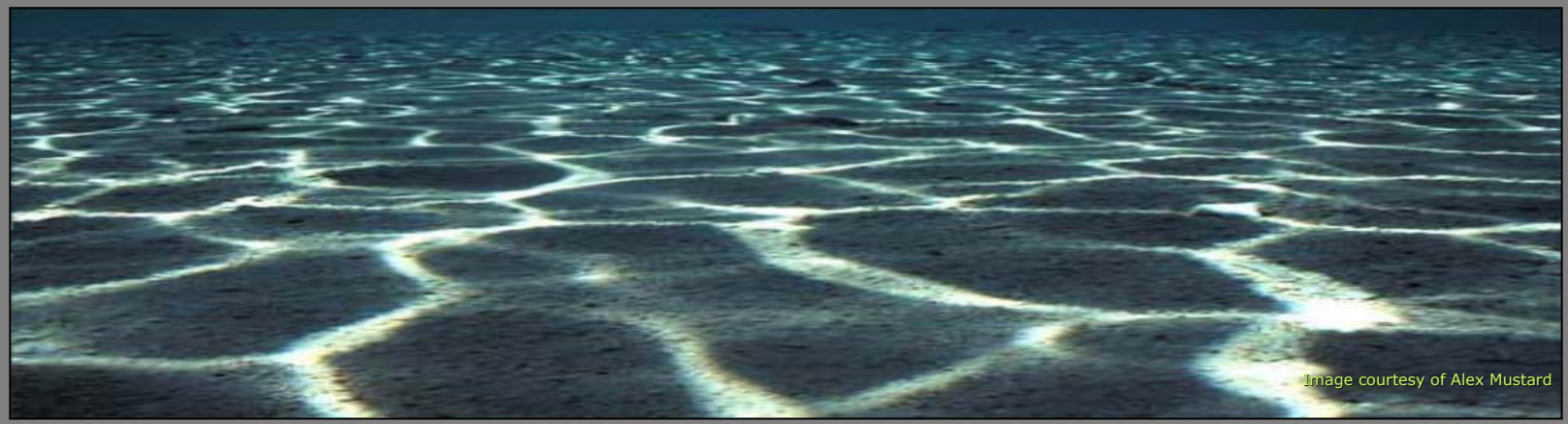
Data Integration Made Easier

Gwenaëlle Moncoiffé

British Oceanographic Data Centre
IOC/IODE GEBICH



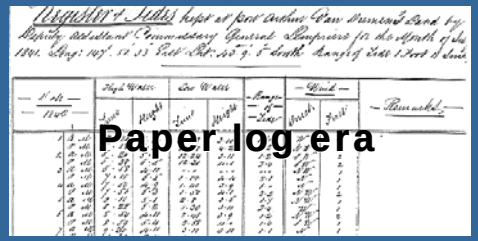
IMBER workshop BEER: Being Efficient and Environmentally Responsible, Miami, 9 November 2008

Introduction

- Data integration: not an option but a necessity.
  - Data integration at the heart of IMBER
  - JGOFS experience ≠ WOCE ≠ GLOBEC> "lessons learnt"
  - Legacy
  - Increasing demand for easily accessible data and for compilation of global datasets for use in climatologies, gridded products
  - E.g. CO2, DMS, data integration activities under SOLAS

No data integration possible without good data management practices

Data Management standards



HIGH

- Strong institute or team-based ethic
- Fewer data better managed
- Long-term staff, stable teams

Personal Computer era

LOW

- Left to the individual
- Demise of the long-term lab technician
- Diversification + short contract + high turnover > degradation of data management standards
- Rules disappeared or were no longer appropriate/adapted
- Data management becomes an IT "thing"

Global networks era

HIGH

- Growing community ethic
- Value in data beyond personal research use
- Need to re-instate strong basic principles for efficient data management
- Re-establish data management as a basic scientific skill & an essential requirement
- Not a technicality but our responsibility as scientists.

Nature vol 455 issue 7209
September 2008

"Researchers need to be obliged to document and manage their data with as much professionalism as they devote to their experiments."

"they should receive greater support in this endeavour than they are afforded at present."

"The lack of standards, for instance, confounds many a researcher seeking to harness the diversity of knowledge now available on any chosen topic."

"All credit, then, to those in the vanguard of interoperability."

"such standards require support from researchers, who should adopt them and deploy them consistently."

- This presentation aims to provide tips on how to make the whole process work more efficiently for individual research scientists and anybody keen to follow good data management practices.
- Practical and non-technical
- Mainly cruise-based examples but could be adapted to any form of data collection activities
- Work in progress which feeds from or will feed into the IMBER Cookbook
- Examples and list of resources will be biased towards areas I am more familiar with: UK/Europe, cruise based data collections;
- Your input to help fill the gaps and fine tune some sections is important.

Rule 1: start early

- Section relevant to all project PIs

- Include data management in your budget! 5-10% of your grant is a good approximation.

- As a project PI / cruise PSO: decide what strategy you will adopt
  - Do it yourself (not recommended even for small projects)- time-consuming and too tempting to prioritise the meaty "publishable" bits
  - Hire somebody to act as a data manager/integrator (full-time or part–time)
  - Delegate task to a [young] scientist embedded in your team or approach your national data centre or institute data management team

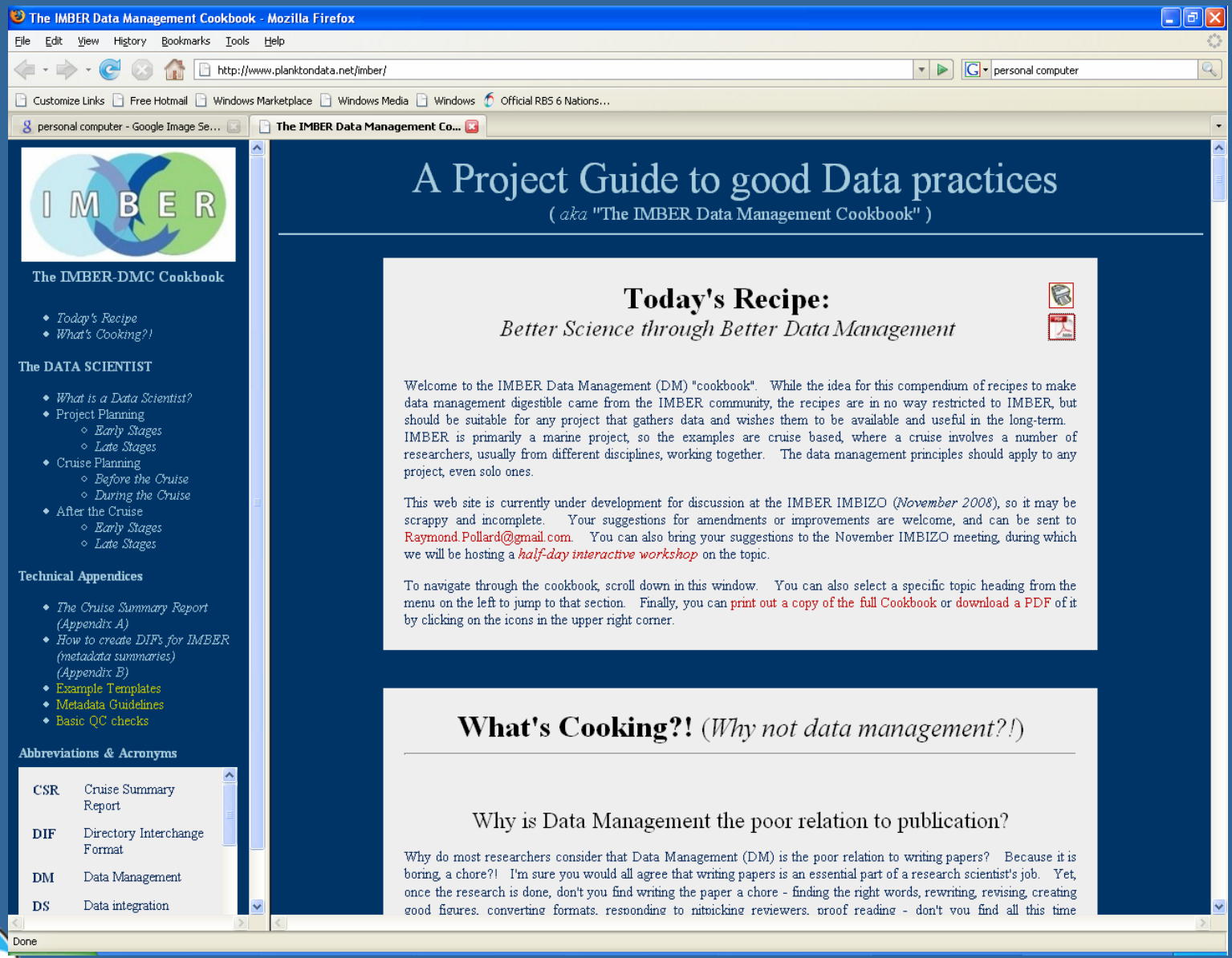
- A data scientist's tasks start before the first cruise starts and end after the last samples have been analysed and the last dataset handed over.

- If multiple projects share the same cruise(s) then money can be pooled to support 1 data management berth and pre- and post-cruise work.

Rule 2: follow the guidelines

- Section relevant to project PIs, data collectors and data scientists

Rule 2: follow the guidelines



- Will facilitate collaboration between IMBER teams, projects, countries or activities ("interoperability")
- Will enable IMBER scientists to make use of existing and developing analytical tools
- Will make it easier to provide data to modellers and global dataset builders
- Will give your data more value by making them more re-usable

Help is at hand: data centres

- NODCs

- Established expert data management centres/teams
  - BCO-DMO
  - CYBER from CNRS at the laboratoire de Villefranche
  - …
- Specialist and world data centres
  - CDIAC, COPEPODS, CCHDO
  - WDC-I Silver Spring, WDC-MARE
  - …

Help is at hand: advising groups, technical support, standard setters

- MMI (Marine Metadata Interoperability) website and very accessible guides on metadata, ontologies, vocabularies
- IODE: Ocean Teacher, Ocean Data Standards, Expert groups
- ICES: Guidelines for data collection and exchange
- CLIVAR and Carbon Hydrographic Data Office (CCHDO): Manuals for data collection and exchange
- GCMD: Discovery metadata, DIFs
- SeaDataNet (in Europe): standards, vocabularies mapping
- …

IMBER: what? where? who? how?

- Span a wide range of measurements: standard core oceanographic measurements, biogeochemistry, biodiversity data, chemical and biological stocks and rates
- From surface to deep ocean sediment
- Research cruises, in situ enrichment experiments, modelling activities, lab and mesocosm experiments, remote sensing

Need a central searchable inventory

Building the IMBER data inventory

- At the Project level
- At the Cruise level
- At Individual level

Minimum requirement

Project information in GCMD

Cruises in CSR databases and eventually DIFs

- GCMD portal but learn from past experience (e.g. GLOBEC)
- Ensure that at least one level of information (the "project" level) is captured consistently and linked to IMBER as a project
- Follow DIF guidelines for IMBER projects (see cookbook)
- Cruise information will be added once a converter CSR to DIF becomes available
- It is good practice to create a DIF for your individual dataset record; in doing this ensure it is linked to IMBER (there will be info in the Cookbook about this)

- Why CSRs?
  - Specifically designed for oceanographic cruise reporting
  - The only existing international standard (initiated by the IOC in the 60s)
  - Two international repositories: DOD and ICES
  - Adopted by European SeaDataNet project and by POGO
  - Online tools are becoming available
  - Long history with an important legacy population (38,000 cruises in ICES database)

Building the IMBER inventory

Integrated Marine Biogeochemistry and Ecosystem Research

Examples of non cruise-based activities:
- Lab and mesocosm experiments
- Remote-sensing
- Socio-economic data
- Model-derived data

Cruises

Filled in by PSOs or appointed project's data specialist

CSRs

Converter to be developed

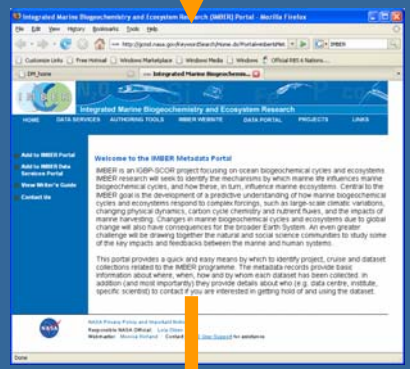
NODC

Projects and non-cruise activities

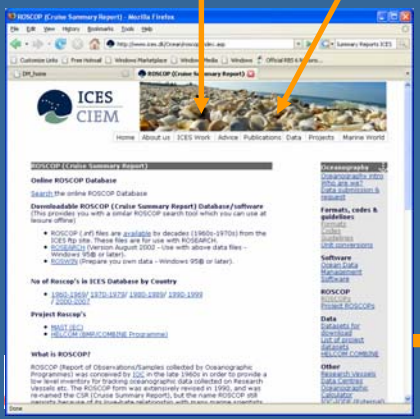
Filled in by projects' data scientists

DIFs

Reviewed by IMBER DLO



IMBER Metadata Portal in GCMD



Data discovery

Integrated Marine Biogeochemistry and Ecosystem Research

Data sharing and archiving

- Section relevant to all data collectors and appointed data scientists

- What to archive?

Three levels	Examples	Archive?	Where?
Raw (digital) data	Plankton raw counts, PVR images, FRRF raw files, etc.	☑☑ YES to enable future reference and for safe keeping.	NODCs, WDCs, digital archives/repositories
Primary (measured) processed	Plankton abundance, biovolume, etc.	☑☑☑ YES, this is the minimum required.	NODCs, digital archives or repositories, national and international SDCs, WDCs
Derived /interpreted data	Plankton biomass using conversion factors, ratio of measured quantities, reduced/combined data, etc.	☑ ok for sharing and sometimes very important to archive but **never** on their own. e.g. Abundance + biovolumes needed alongside plankton biomass	NODCs, digital archives or repositories, national and international SDCs, WDCs

- How to share and archive data?
- References
  - Best Practices for Preparing Environmental Data Sets to Share and Archive (Cook et al 2001, updated by Hook et al 2007, Oak Ridge National Laboratory) – 9 pages.
    - http//daac.ornl.gov/PI/bestprac.html
  - BCO-DMO Data Management Guidelines Manual: *a collection of best practice recommendations for collecting and sharing biogeochemical and ecological oceanographic data and metadata – 15 pages.*
    - http://bcodmo.org/files/bcodmo/BCO-DMO_best_prac_v1d1.pdf

- The 7 best practices for environmental data (http@//daac.ornl.gov/PI/bestprac.html)
  1. Assign descriptive (and unique) file names
  2. Use consistent and stable file formats
  3. Define the content of your data files
  4. Use consistent data organisation
  5. Perform basic quality assurance
  6. Assign descriptive data set titles
  7. Provide documentation

- 1. Assign descriptive file names
  - Reflect content and uniquely identify the data file
  - Use project acronym, cruise identifier, station identifier, study title, investigator, data type, version number, date created, etc.
  - Compliant with various data systems and platforms
    - only numbers, letter, dashes, underscores – no space or special characters
    - Use lower case
    - Limit length to 64 characters max
    - Use appropriate file extension to reflect file type.

D999_ctd001_pro_20080901.cnv

D999_frrf001_pro_20080901.asc

D999_frrf001_raw.asc

D999_uway_nutrients_20080901.xls

D999_ctd_nutrients_20080901.xls

D999_blogg_hplc_v1_20080901.csv

D999_blogg_hplc_bodc_20090901.csv

D999_blogg_hplc_bodc_20081201.csv

Bodc.xls

Data for bodc.xls

Nutrients.csv

Hplc.csv

BODC

NATURAL ENVIRONMENT RESEARCH COUNCIL

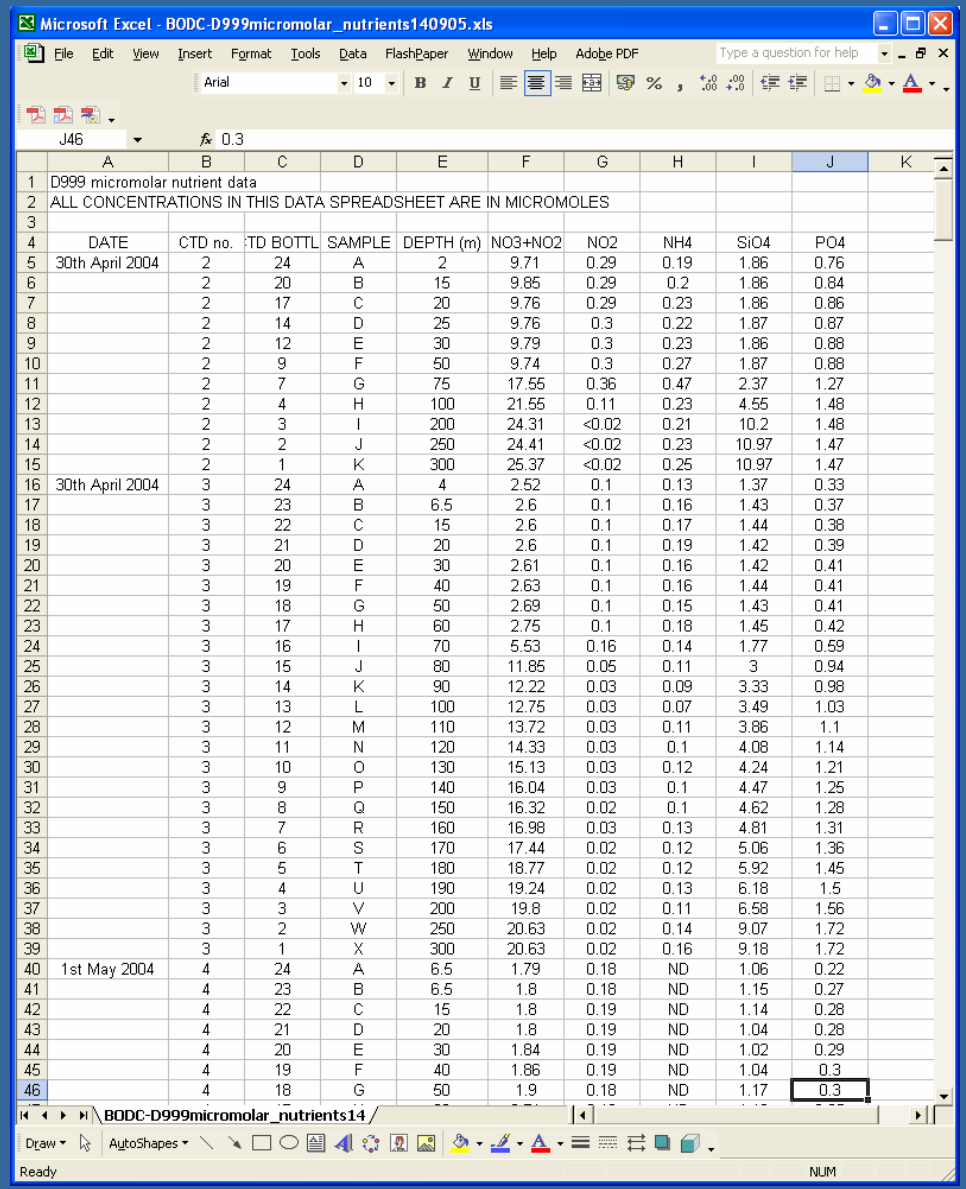
IMBER

Integrated Marine Biogeochemistry and Ecosystem Research

- 2. Use consistent and stable file formats
  - Tabular data
  - Spreadsheet and ASCII format
  - Use the same format throughout the file or files
  - For ASCII, use delimiters such as comma, tab or semicolon
  - Don't include figures and summary statistics in the data file – keep these in a separate file
  - Use a header
    - Avoid special characters and avoid using the chosen delimiter
    - Identify your dataset with data file name, data set title, author, version name, date created, date modified, reason for modifications, associated document file name
    - Use column headings for parameter names
    - Use one row for parameter units

- 3. Define the content of your data files
  - Parameter Name
    - Use commonly accepted parameter names
    - Always define abbreviated parameter names in data documentation
    - Use consistent capitalisation
    - No special character and use underscores to replace spaces
  - Units
    - Explicit in data file and in documentation
    - Use recommended units as much as possible (see for example http://ijgofs.whoi.edu/D_I_M/core_parameters_Dec2003_final.pdf)
    - If expressing concentrations in units "per kg-1" (WOCE standard http://whpo.ucsd.edu/manuals/pdf/90_1/appendxg.pdf) include the information necessary to convert the value on a unit "per litre".

Spreadsheet example 1



Microsoft Excel - BODC-D999micromolar_nutrients140905.xls

File Edit View Insert Format Tools Data FlashPaper Window Help Adobe PDF

Arial 10

J46 0.3

	A	B	C	D	E	F	G	H	I	J	K
1	D999 micromolar nutrient data										
2	ALL CONCENTRATIONS IN THIS DATA SPREADSHEET ARE IN MICROMOLES										
3											
4	DATE	CTD no.	TD BOTTL	SAMPLE	DEPTH (m)	NO3+NO2	NO2	NH4	SiO4	PO4	
5	30th April 2004	2	24	A	2	9.71	0.29	0.19	1.86	0.76	
6		2	20	B	15	9.85	0.29	0.2	1.86	0.84	
7		2	17	C	20	9.76	0.29	0.23	1.86	0.86	
8		2	14	D	25	9.76	0.3	0.22	1.87	0.87	
9		2	12	E	30	9.79	0.3	0.23	1.86	0.88	
10		2	9	F	50	9.74	0.3	0.27	1.87	0.88	
11		2	7	G	75	17.55	0.36	0.47	2.37	1.27	
12		2	4	H	100	21.55	0.11	0.23	4.55	1.48	
13		2	3	I	200	24.31	<0.02	0.21	10.2	1.48	
14		2	2	J	250	24.41	<0.02	0.23	10.97	1.47	
15		2	1	K	300	25.37	<0.02	0.25	10.97	1.47	
16	30th April 2004	3	24	A	4	2.52	0.1	0.13	1.37	0.33	
17		3	23	B	6.5	2.6	0.1	0.16	1.43	0.37	
18		3	22	C	15	2.6	0.1	0.17	1.44	0.38	
19		3	21	D	20	2.6	0.1	0.19	1.42	0.39	
20		3	20	E	30	2.61	0.1	0.16	1.42	0.41	
21		3	19	F	40	2.63	0.1	0.16	1.44	0.41	
22		3	18	G	50	2.69	0.1	0.15	1.43	0.41	
23		3	17	H	60	2.75	0.1	0.18	1.45	0.42	
24		3	16	I	70	5.53	0.16	0.14	1.77	0.59	
25		3	15	J	80	11.85	0.05	0.11	3	0.94	
26		3	14	K	90	12.22	0.03	0.09	3.33	0.98	
27		3	13	L	100	12.75	0.03	0.07	3.49	1.03	
28		3	12	M	110	13.72	0.03	0.11	3.86	1.1	
29		3	11	N	120	14.33	0.03	0.1	4.08	1.14	
30		3	10	O	130	15.13	0.03	0.12	4.24	1.21	
31		3	9	P	140	16.04	0.03	0.1	4.47	1.25	
32		3	8	Q	150	16.32	0.02	0.1	4.62	1.28	
33		3	7	R	160	16.98	0.03	0.13	4.81	1.31	
34		3	6	S	170	17.44	0.02	0.12	5.06	1.36	
35		3	5	T	180	18.77	0.02	0.12	5.92	1.45	
36		3	4	U	190	19.24	0.02	0.13	6.18	1.5	
37		3	3	V	200	19.8	0.02	0.11	6.58	1.56	
38		3	2	W	250	20.63	0.02	0.14	9.07	1.72	
39		3	1	X	300	20.63	0.02	0.16	9.18	1.72	
40	1st May 2004	4	24	A	6.5	1.79	0.18	ND	1.06	0.22	
41		4	23	B	6.5	1.8	0.18	ND	1.15	0.27	
42		4	22	C	15	1.8	0.19	ND	1.14	0.28	
43		4	21	D	20	1.8	0.19	ND	1.04	0.28	
44		4	20	E	30	1.84	0.19	ND	1.02	0.29	
45		4	19	F	40	1.86	0.19	ND	1.04	0.3	
46		4	18	G	50	1.9	0.18	ND	1.17	0.3	

BODC-D999micromolar_nutrients14

Draw AutoShapes

Ready NUM

Examples of data spreadsheet submission to the BODC

BODC has an history of not being prescriptive with file format – issues list of requirements and preferences

Advantage: scientists do not have to reformat their data specially for us

Disadvantage: more time-consuming to ingest; variation in data layout seems to be unlimited; and scientists easily forget our recommendations.

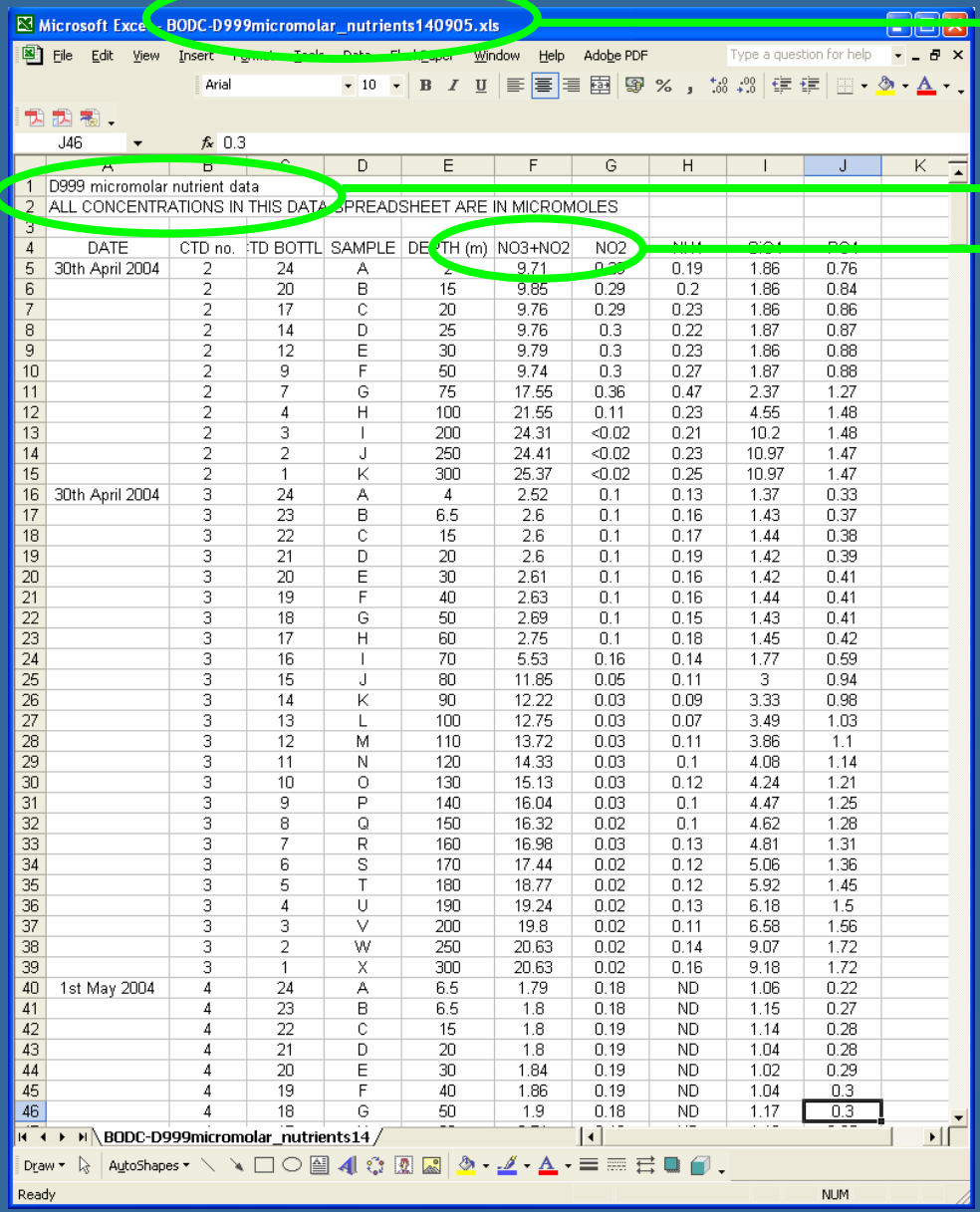
Solution? Distribute clearer guidelines (checklist?)? Provide examples?

BODC

NATURAL ENVIRONMENT RESEARCH COUNCIL

IMBER Integrated Marine Biogeochemistry and Ecosystem Research

Spreadsheet example 1



Good filename

Good data set title

Good explicit column headers

✓ Good overall data organisation

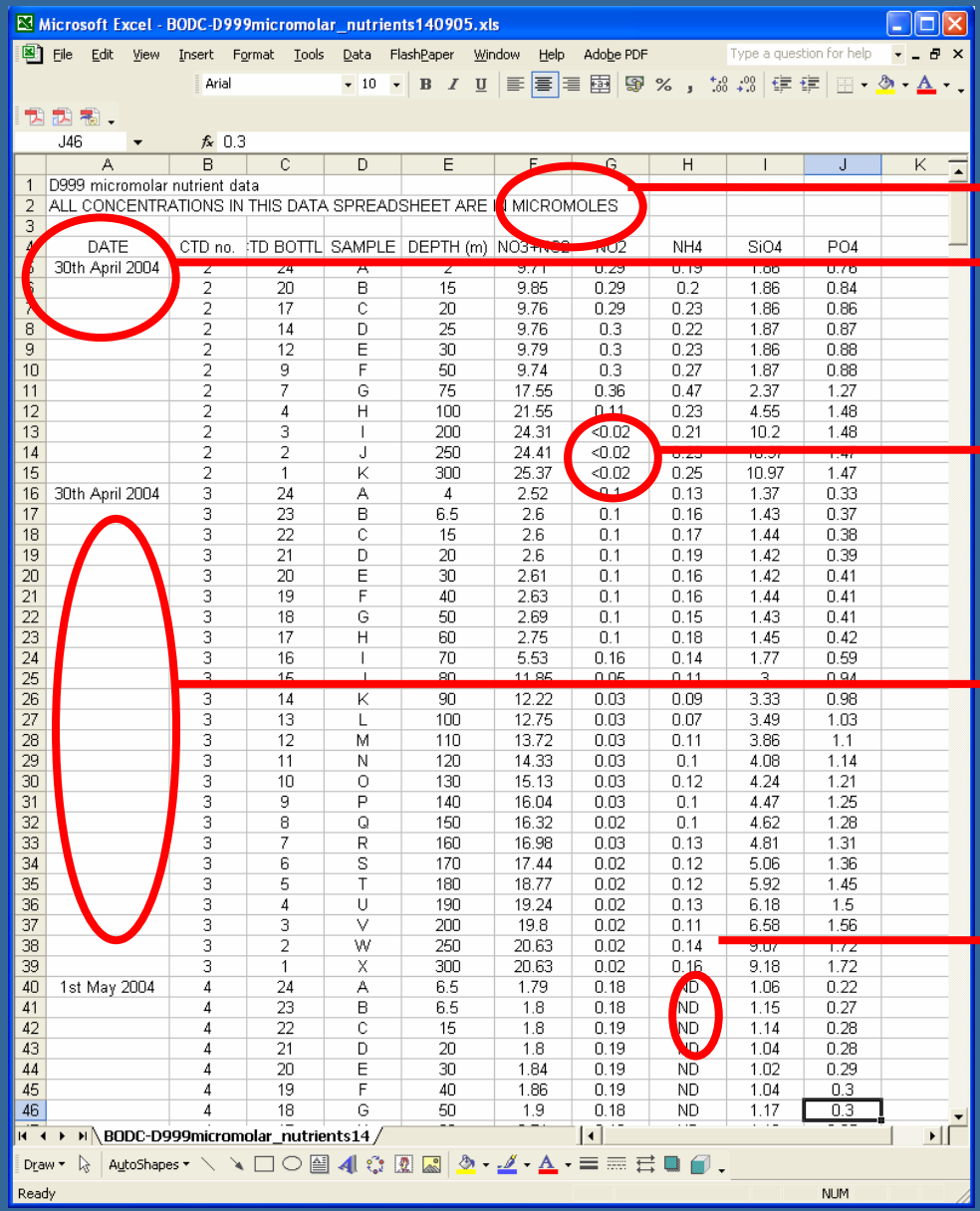
BODC

NATURAL ENVIRONMENT RESEARCH COUNCIL

IMBER

Integrated Marine Biogeochemistry and Ecosystem Research

Spreadsheet example 1



Definitions need to be explicit.

Avoid free text for dates - use standard machine readable date and time formats.

Do not mix characters and numbers in same field. Use separate columns for flags.

Avoid blank cells unless the value is missing.

Use consistent data format AND indicate the convention used for absent values in the header.

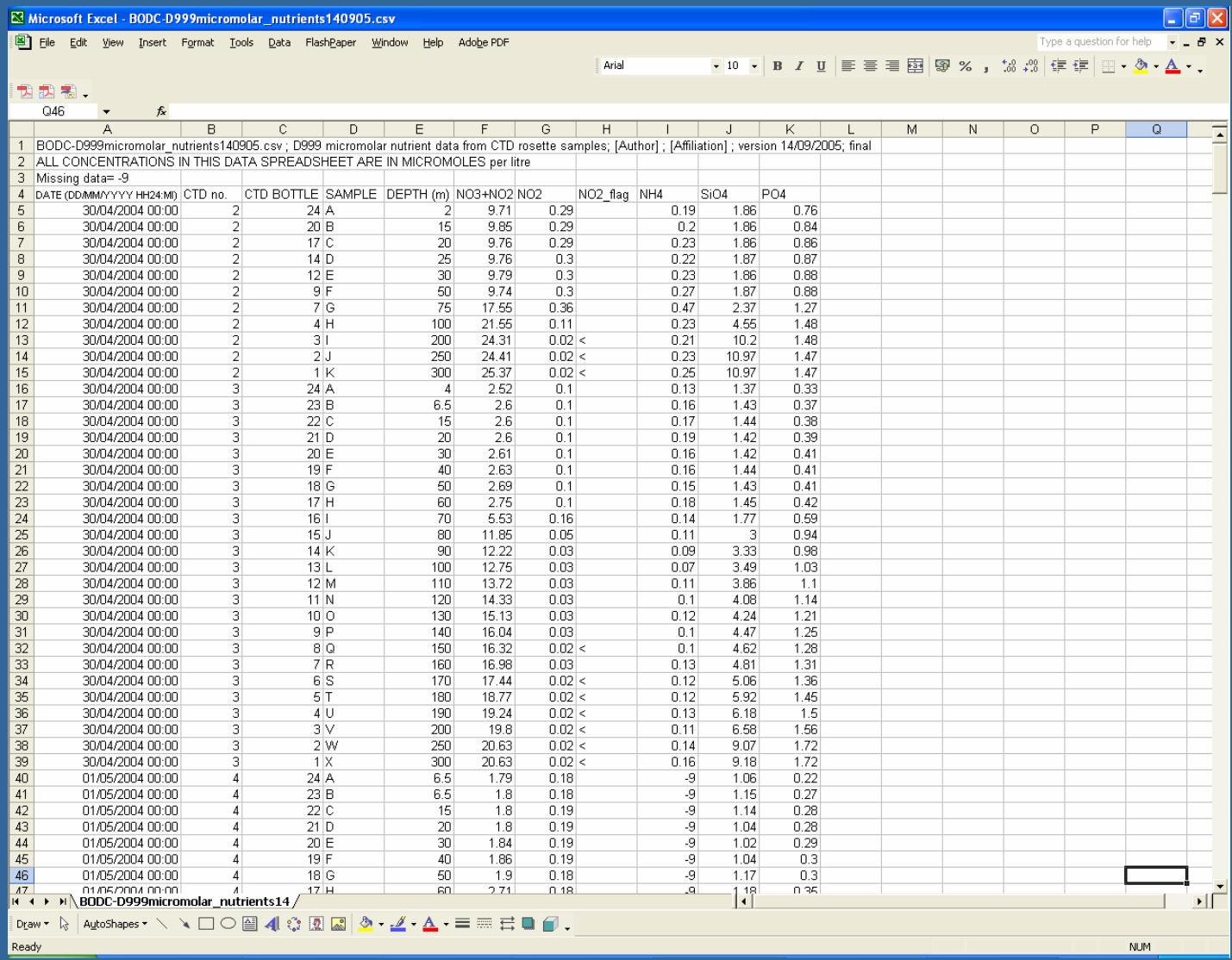
BODC

NATURAL ENVIRONMENT RESEARCH COUNCIL

IMBER Integrated Marine Biogeochemistry and Ecosystem Research

Spreadsheet example 1

Example of a data file formatted for easy ingestion in any data system



Microsoft Excel - BODC-D999micromolar_nutrients140905.csv

	A	B	C	D	E	F	G	H	I	J	K
1	BODC-D999micromolar_nutrients140905.csv ; D999 micromolar nutrient data from CTD rosette samples; [Author] ; [Affiliation] ; version 14/09/2005; final										
2	ALL CONCENTRATIONS IN THIS DATA SPREADSHEET ARE IN MICROMOLES per litre										
3	Missing data= -9										
4	DATE (DD/MM/YYYY HH24:MI)	CTD no.	CTD BOTTLE	SAMPLE	DEPTH (m)	NO3+NO2	NO2	NO2_flag	NH4	SiO4	PO4
5	30/04/2004 00:00	2	24	A	2	9.71	0.29		0.19	1.86	0.76
6	30/04/2004 00:00	2	20	B	15	9.85	0.29		0.2	1.86	0.84
7	30/04/2004 00:00	2	17	C	20	9.76	0.29		0.23	1.86	0.86
8	30/04/2004 00:00	2	14	D	25	9.76	0.3		0.22	1.87	0.87
9	30/04/2004 00:00	2	12	E	30	9.79	0.3		0.23	1.86	0.88
10	30/04/2004 00:00	2	9	F	50	9.74	0.3		0.27	1.87	0.88
11	30/04/2004 00:00	2	7	G	75	17.55	0.36		0.47	2.37	1.27
12	30/04/2004 00:00	2	4	H	100	21.55	0.11		0.23	4.55	1.48
13	30/04/2004 00:00	2	3	I	200	24.31	0.02	<	0.21	10.2	1.48
14	30/04/2004 00:00	2	2	J	250	24.41	0.02	<	0.23	10.97	1.47
15	30/04/2004 00:00	2	1	K	300	25.37	0.02	<	0.25	10.97	1.47
16	30/04/2004 00:00	3	24	A	4	2.52	0.1		0.13	1.37	0.33
17	30/04/2004 00:00	3	23	B	6.5	2.6	0.1		0.16	1.43	0.37
18	30/04/2004 00:00	3	22	C	15	2.6	0.1		0.17	1.44	0.38
19	30/04/2004 00:00	3	21	D	20	2.6	0.1		0.19	1.42	0.39
20	30/04/2004 00:00	3	20	E	30	2.61	0.1		0.16	1.42	0.41
21	30/04/2004 00:00	3	19	F	40	2.63	0.1		0.16	1.44	0.41
22	30/04/2004 00:00	3	18	G	50	2.69	0.1		0.15	1.43	0.41
23	30/04/2004 00:00	3	17	H	60	2.75	0.1		0.18	1.45	0.42
24	30/04/2004 00:00	3	16	I	70	5.53	0.16		0.14	1.77	0.59
25	30/04/2004 00:00	3	15	J	80	11.85	0.05		0.11	3	0.94
26	30/04/2004 00:00	3	14	K	90	12.22	0.03		0.09	3.33	0.98
27	30/04/2004 00:00	3	13	L	100	12.75	0.03		0.07	3.49	1.03
28	30/04/2004 00:00	3	12	M	110	13.72	0.03		0.11	3.86	1.1
29	30/04/2004 00:00	3	11	N	120	14.33	0.03		0.1	4.08	1.14
30	30/04/2004 00:00	3	10	O	130	15.13	0.03		0.12	4.24	1.21
31	30/04/2004 00:00	3	9	P	140	16.04	0.03		0.1	4.47	1.25
32	30/04/2004 00:00	3	8	Q	150	16.32	0.02	<	0.1	4.62	1.28
33	30/04/2004 00:00	3	7	R	160	16.98	0.03		0.13	4.81	1.31
34	30/04/2004 00:00	3	6	S	170	17.44	0.02	<	0.12	5.06	1.36
35	30/04/2004 00:00	3	5	T	180	18.77	0.02	<	0.12	5.92	1.45
36	30/04/2004 00:00	3	4	U	190	19.24	0.02	<	0.13	6.18	1.5
37	30/04/2004 00:00	3	3	V	200	19.8	0.02	<	0.11	6.58	1.56
38	30/04/2004 00:00	3	2	W	250	20.63	0.02	<	0.14	9.07	1.72
39	30/04/2004 00:00	3	1	X	300	20.63	0.02	<	0.16	9.18	1.72
40	01/05/2004 00:00	4	24	A	6.5	1.79	0.18		-9	1.06	0.22
41	01/05/2004 00:00	4	23	B	6.5	1.8	0.18		-9	1.15	0.27
42	01/05/2004 00:00	4	22	C	15	1.8	0.19		-9	1.14	0.28
43	01/05/2004 00:00	4	21	D	20	1.8	0.19		-9	1.04	0.28
44	01/05/2004 00:00	4	20	E	30	1.84	0.19		-9	1.02	0.29
45	01/05/2004 00:00	4	19	F	40	1.86	0.19		-9	1.04	0.3
46	01/05/2004 00:00	4	18	G	50	1.9	0.18		-9	1.17	0.3
47	01/05/2004 00:00	4	17	H	60	2.71	0.18		-9	1.18	0.35

BODC-D999micromolar_nutrients14

Ready

NUM

Integrated Marine Biogeochemistry and Ecosystem Research

Spreadsheet example 2

Original data file

Microsoft Excel - in-situ O2 for BODC.xls

	A	B	C	D	E	F	G
3	CTD #	1		pre dawn			
5	Oxygen #	depth	Niskin	fix temp	sal	temp	umol/l winkler
6	942	4	24	16.6	35.4	18.9	236.12
7	943	12	20	17.2	35.4	18.9	235.80
8	945	32	14	15.7	35.4	18.8	232.92
9	946	52	11	17.1	35.4	18.7	230.53
10	947	86	6	14.6	35.2	13.2	212.36
11	948	102	5	13.8	35.1	12.3	213.81
12	949	127	4	13	34.9	11.2	194.94
13	898	178	3	12	34.8	10	177.36
14	944	254	1	10.6	34.6	7.9	157.74
16	ctd #	2		optics			
18	Oxygen #	depth	Niskin	fix temp	sal	temp	umol/l winkler
19	741	2	24	17.3	35.2	16.8	
20	742	10	22	17	35.2	16.7	252.35
21	743	11	20	17	35.2	16.7	252.53
22	744	31	15	16.3	35.2	16	239.71
23	745	50	14	14.6	35.2	13.4	218.25
24	746	102	8	13.7	35.1	12.3	218.52
25	194	151	5	12.6	34.9	11	213.27
26	195	252	3	11	34.7	8.9	184.95

CTD 210505 | CTD 240505 | CTD 250505 | CTD 260505 | CTD 270505 | CTD 280505 | CTD 290505 | CTD 300505 | CTD 310

Ready

Well organised data file but extremely difficult to ingest in a database because split in multiple worksheets

BODC

NATURAL ENVIRONMENT RESEARCH COUNCIL

IMBER Integrated Marine Biogeochemistry and Ecosystem Research

Spreadsheet example 2

Reformatted data file

Microsoft Excel - in-situ_O2_for_BODC_for_banking.xls

File Edit View Insert Format Tools Data FlashPaper Window Help Adobe PDF

Arial 10

G21 35.6109

	A	B	C	D	E	F	G	H	I	J
1	cruise	station ID	Oxygen #	depth	Niskin	fix temp	CTD sal	CTD temp	umol/l winkler	
2	D999	CTD001	942	4	24	16.6	35.4	18.9	236.12	
3	D999	CTD001	943	12	20	17.2	35.4	18.9	235.80	
4	D999	CTD001	945	32	14	15.7	35.4	18.8	232.92	
5	D999	CTD001	946	52	11	17.1	35.4	18.7	230.53	
6	D999	CTD001	947	86	6	14.6	35.2	13.2	212.36	
7	D999	CTD001	948	102	5	13.8	35.1	12.3	213.81	
8	D999	CTD001	949	127	4	13.0	34.9	11.2	194.94	
9	D999	CTD001	898	178	3	12.0	34.8	10.0	177.36	
10	D999	CTD001	944	254	1	10.6	34.6	7.9	157.74	
11	D999	CTD002	741	2	24	17.3	35.2	16.8		
12	D999	CTD002	742	10	22	17.0	35.2	16.7	252.35	
13	D999	CTD002	743	11	20	17.0	35.2	16.7	252.53	
14	D999	CTD002	744	31	15	16.3	35.2	16.0	239.71	
15	D999	CTD002	745	50	14	14.6	35.2	13.4	218.25	
16	D999	CTD002	746	102	8	13.7	35.1	12.3	218.52	
17	D999	CTD002	194	151	5	12.6	34.9	11.0	213.27	
18	D999	CTD002	195	252	3	11.0	34.7	8.9	184.95	
19	D999	CTD003	942	16	20	17.1	35.5	19.2	236.27	
20	D999	CTD003	943	27	14	16.8	35.6	19.0	241.59	
21	D999	CTD003	944	46	12	17.1	35.6	19.0	237.07	
22	D999	CTD003	945	100	9	17.0	35.6	18.7	231.58	
23	D999	CTD003	946	148	4	16.8	35.5	16.4	231.51	
24	D999	CTD003	947	207	3	15.7	35.5	15.7	228.29	
25	D999	CTD003	948	254	2	15.1	35.3	14.6	221.59	
26	D999	CTD003	949	308	1	14.8	35.3	13.9		
27	D999	CTD004	194	2.847	23	19.5	35.7	19.6	234.97	
28	D999	CTD004	195	14.253	21	19.1	35.7	19.6	234.90	
29	D999	CTD004	196	24.56	19	18.9	35.7	19.6	234.81	
30	D999	CTD004	197	43.359	17	18.9	35.7	19.6	234.64	
31	D999	CTD004	198	90.656	15	18.8	35.7	19.6	234.58	
32	D999	CTD004	199	95.348	14	18.9	35.7	19.6	234.40	
33	D999	CTD004	244	100.532	10	18.9	35.7	19.6	234.41	
34	D999	CTD004	245	110.248	9	18.8	35.7	19.5	235.94	
35	D999	CTD004	246	125.573	6	17.8	35.5	17.3	214.08	
36	D999	CTD004	247	150.028	5	16.9	35.4	15.6	208.88	
37	D999	CTD004	249	301.527	1	14.6	35.1	12.2	220.62	

for_loading

Draw AutoShapes

Ready NUM

BODC

NATURAL ENVIRONMENT RESEARCH COUNCIL

IMBER Integrated Marine Biogeochemistry and Ecosystem Research

Other issues to look for

- Depth poorly defined: e.g. “SSCM”, “1%LD”, “bottom”, “0m”, “mixed layer”, “O2 minimum”
- Underway samples identified by their position only (no date and time)
- CTD samples identified by a “site” name only although multiple CTD casts were done at that site at different times of day (and night).
- Sampling identified by a date only: no station identifiers, no time
- Use of coloured cells for quality indicators: not machine readable and not safe for long-term preservation (not preserved in ASCII format)

- 4. Use consistent data organisation
  - matrix layout

  - Keep similar data together
  - Avoid very large files **BUT**
  - Do not break up the data in many small data files or multiple worksheets.
  - Only split files when necessary on ground of file size or different metadata field requirements.

- 5. Perform basic quality assurance
  - Check file format in ASCII version
  - Check file organisation and descriptors
  - Ensure all essential metadata is included
  - Check that all the values have transferred correctly to the ASCII version.
  - In particular, check for missing reference indicator "#REF!"
  - Convert formula to value before saving CSV file.

- 6. Assign descriptive data set titles

  - Data set titles should contain information about the type of data and for example:
    - the date range, the location, and/or the instruments used;
    - If your data set is part of a larger field project, the project name too.
  - Keep title length to less than 80 characters (spaces included) as much as possible
  - Names should contain only numbers, letters, dashes, underscores and spaces -- no special characters.
  - The data set title should be consistent with the name(s) of the data file(s) in the data archive whether the data set contain only one data file or many thousands data files.

Think Scientific Publication

- 7. Provide data set documentation

  - Title = data set title
  - Author(s), affiliation and contact person
  - Background: reason for collecting the data, funding source and project/programme.
  - Material and methods: sampling strategy and methodology, analytical procedures, quality control procedures
  - Data set content description
  - Short quality assessment to report problems that may limit the use of the data
  - References

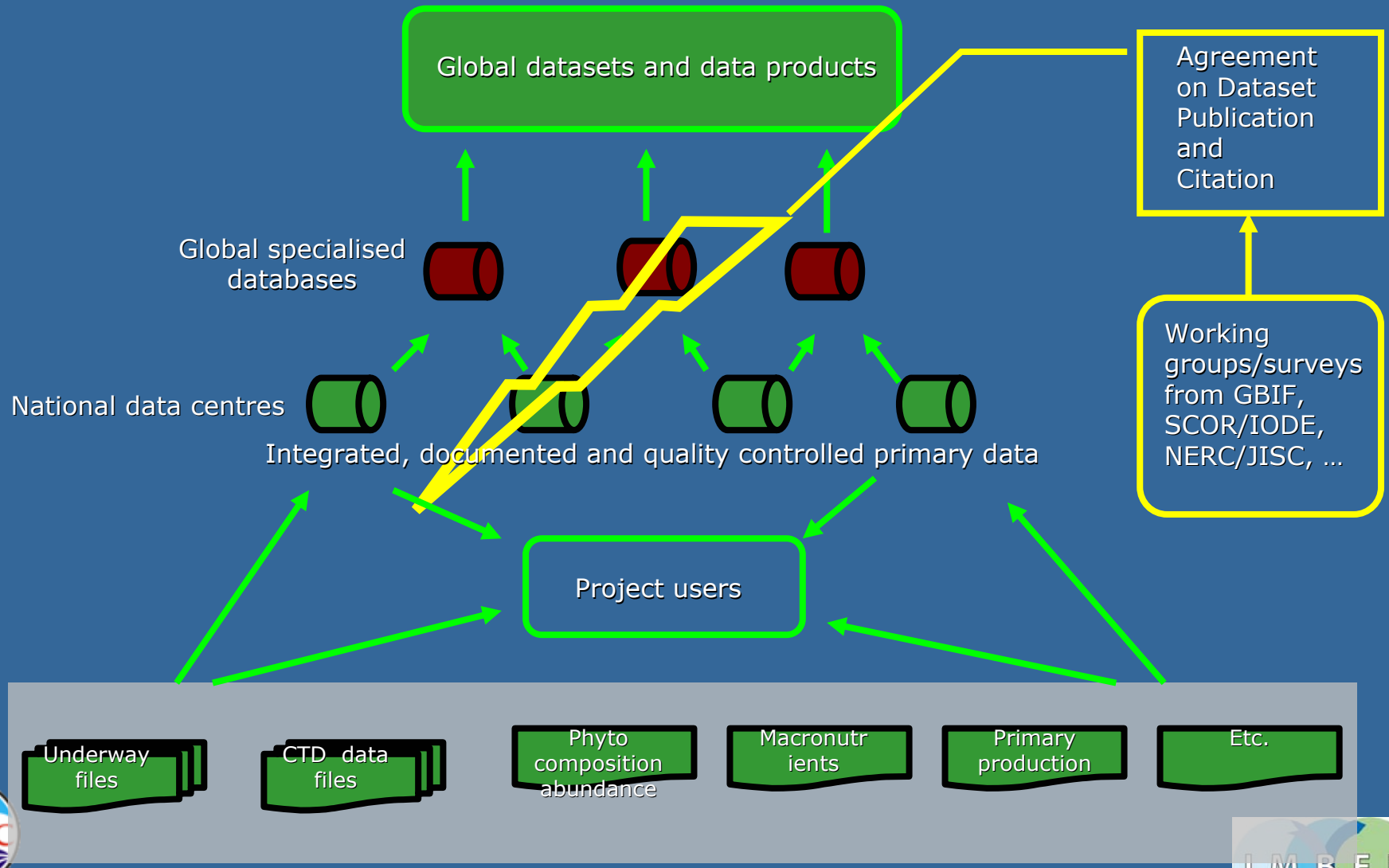
Archiving the data files

- Section relevant to all data collectors and appointed data scientists

Archiving the data files

- Make your master files read-only
- Take copies on a range of media
- Send a copy to your national data centre with instructions on access policy if not already agreed
- Create a DIF for your dataset in GCMD
- Archive as soon as possible and as soon as you start sharing the data – use version numbers and documentation to clearly identify the version of the data

The environmentally responsible data flow



BODC

NATURAL ENVIRONMENT RESEARCH COUNCIL

IMBER

Integrated Marine Biogeochemistry and Ecosystem Research