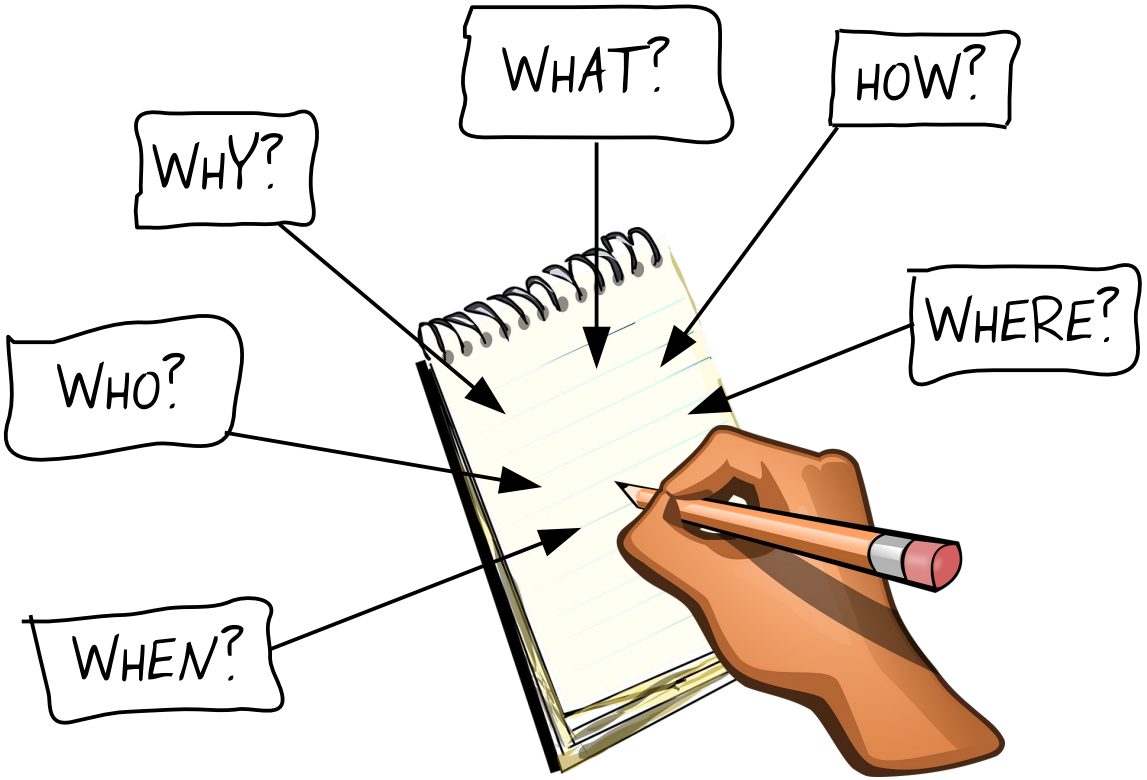
Data acquisition
and impacts on data quality

Sarah Faulwetter

A Practical introduction to acquisition, validation, quality control and access to (biodiversity) data

Vlissingen, Netherlands, 14.06.2015

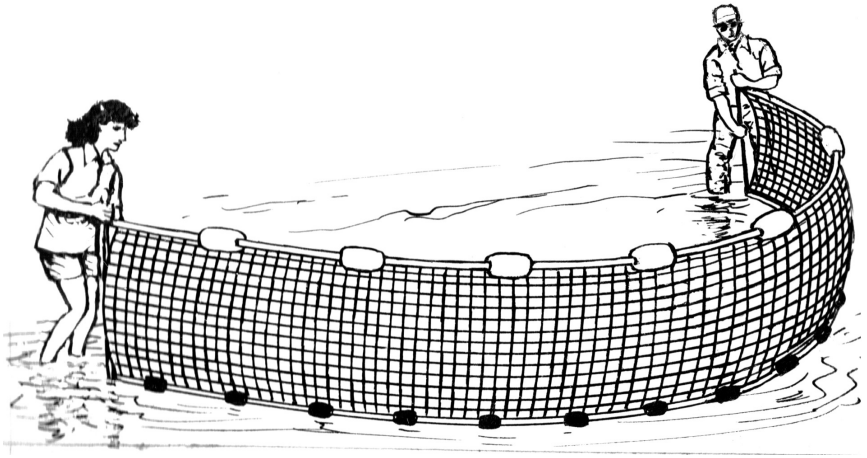
Make a plan



Make a plan

Make a sampling plan

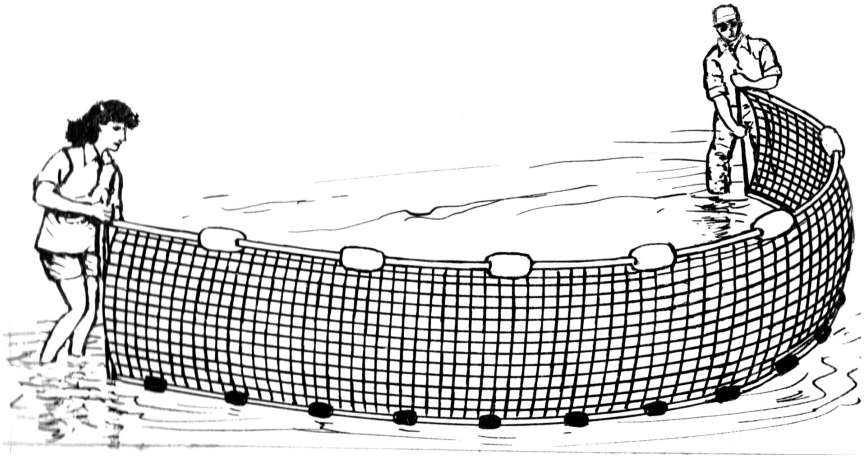
Make a data management plan

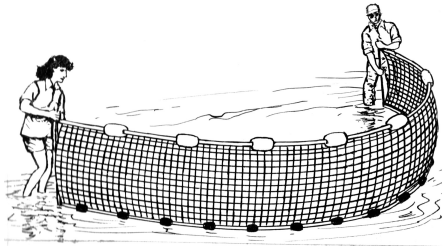


Make a plan

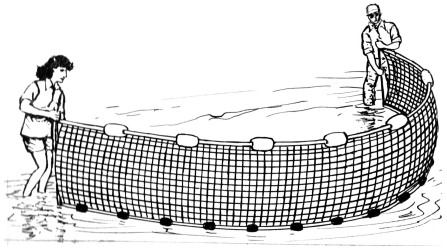
Make a sampling plan

Make a data management plan





WHY?



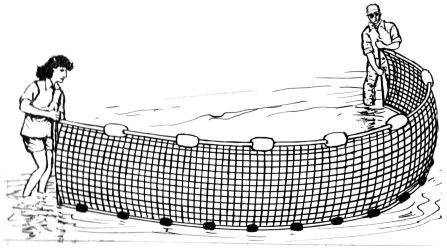
WHY?

Aim of the study?

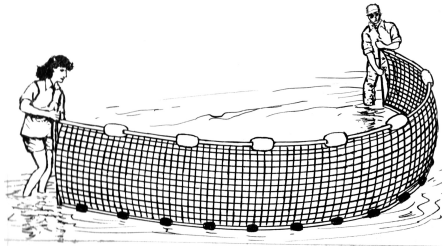
Planned analyses?

Data isolated or to be integrated?

Interoperability?

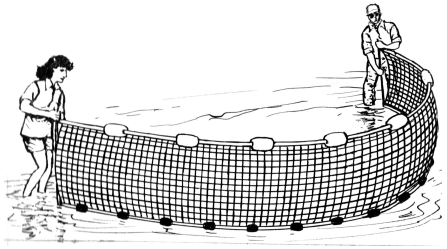


HOW ?



protocol &
templates

standards,
abbreviations,
units &
codes

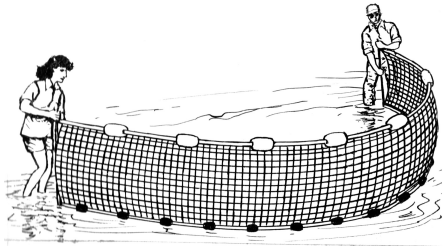


Standardize workflow
Don't forget steps
Recap what has been done

HOW ?

protocol &
templates

standards,
abbreviations,
units &
codes



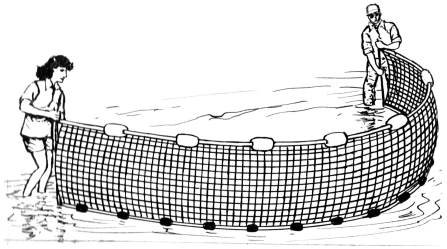
Standardize workflow
Don't forget steps
Recap what has been done

protocol &
templates

HOW ?

WHAT?

standards,
abbreviations,
units &
codes



Standardize workflow
Don't forget steps
Recap what has been done

protocol &
templates

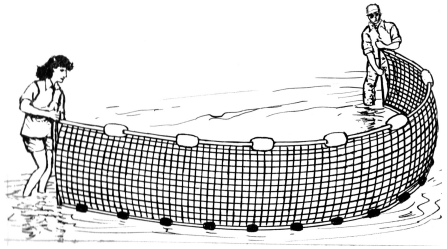
HOW ?

standards,
abbreviations,
units &
codes

data

WHAT?

metadata



Standardize workflow
Don't forget steps
Recap what has been done

protocol &
templates

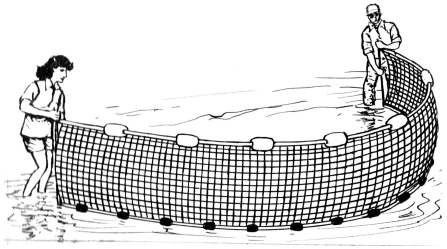
HOW ?

standards,
abbreviations,
units &
codes

data

who, when, how, why, where
problems, notes, remarks
everything that affects data quality

metadata



Standardize workflow
Don't forget steps
Recap what has been done

protocol &
templates

HOW ?

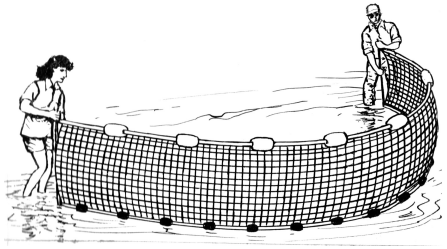
standards,
abbreviations,
units &
codes

data

who, when, how, why, where
problems, notes, remarks
everything that affects data quality

metadata

WHEN?



Standardize workflow
Don't forget steps
Recap what has been done

protocol &
templates

HOW ?

standards,
abbreviations,
units &
codes

data

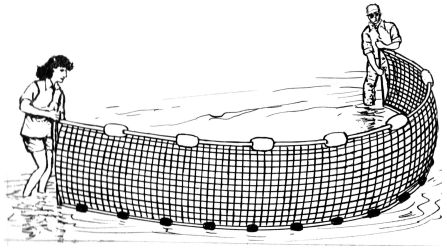
who, when, how, why, where
problems, notes, remarks
everything that affects data quality

metadata

one-time or
continuous?

WHEN?

data entry &
data checks



Standardize workflow
Don't forget steps
Recap what has been done

HOW ?

protocol &
templates

standards,
abbreviations,
units &
codes

data

who, when, how, why, where
problems, notes, remarks
everything that affects data quality

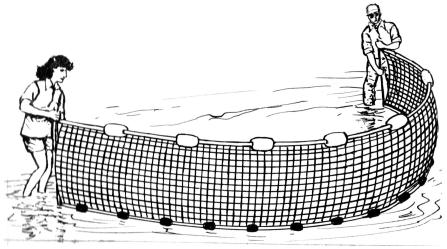
metadata

Fix time points
for tasks!

WHEN?

one-time or
continuous?

data entry &
data checks



Standardize workflow
Don't forget steps
Recap what has been done

protocol &
templates

HOW ?

standards,
abbreviations,
units &
codes

data

who, when, how, why, where
problems, notes, remarks
everything that affects data quality

metadata

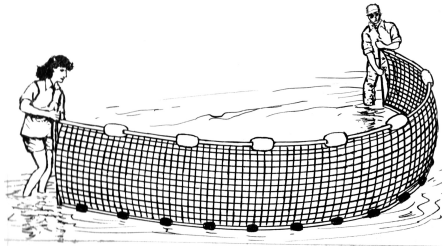
Fix time points
for tasks!

one-time or
continuous?

WHEN?

data entry &
data checks

PLAN FOR
PROBLEMS



Standardize workflow
Don't forget steps
Recap what has been done

protocol & templates

HOW ?

standards, abbreviations, units & codes

data

who, when, how, why, where
problems, notes, remarks
everything that affects data quality

metadata

Fix time points for tasks!

one-time or continuous?

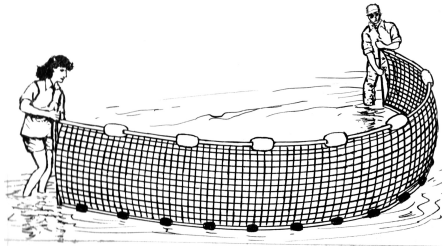
WHEN?

data entry & data checks

which factors influence data quality?

PLAN FOR PROBLEMS

completeness, consistency, validity, accuracy



Standardize workflow
Don't forget steps
Recap what has been done

HOW ?

protocol & templates

standards, abbreviations, units & codes

data

who, when, how, why, where
problems, notes, remarks
everything that affects data quality

metadata

one-time or continuous?

WHEN?

Fix time points for tasks!

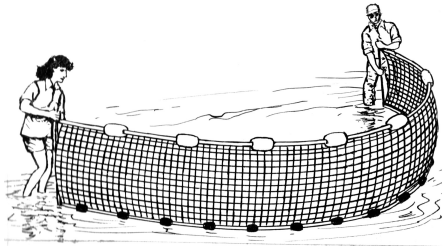
data entry & data checks

Can problems be prevented / migitated?
If not, document them!

PROBLEMS

which factors influence data quality?

completeness, consistency, validity, accuracy



Standardize workflow
Don't forget steps
Recap what has been done

protocol &
templates

HOW ?

standards,
abbreviations,
units &
codes

data

who, when, how, why
problems, notes, remarks
everything that affects data quality

metadata

Fix time points
for tasks!

one-time or
continuous?

WHEN?

data entry &
data checks

Can problems be prevented / migitated?
If not, document them!

PROBLEM

which factors
influence data
quality?

completeness,
consistency,
validity,
accuracy

Aim: data re-usable
by a third person

Aim: data re-usable
by a third person

Documentation!

Aim: data re-usable by a third person

Good documentation = at dataset level AND record level

Documentation!

Aim: data re-usable
by a third person

Good documentation =
at dataset level AND
record level

Documentation!

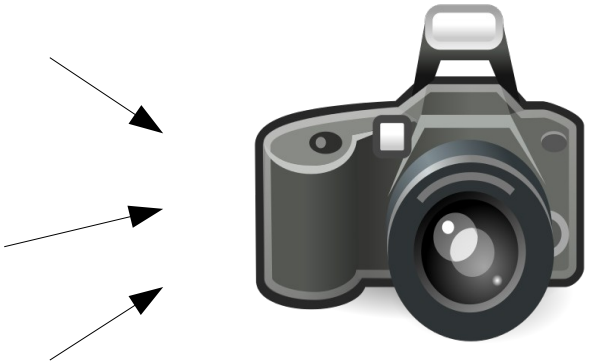
- keep raw data and metadata
- take generous notes & remarks
- note deviations from protocol
- note type of instruments / calibrations
- record literature used (e.g. keys)
- record persons involved with each step
- record data conversions, modifications, validations...
- record why certain steps were taken

Aim: data re-usable
by a third person

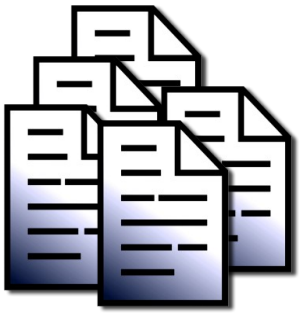
Good documentation =
at dataset level AND
record level

Documentation!

- keep raw data and metadata
- take generous notes & remarks
- note deviations from protocol
- note type of instruments / calibrations
- record literature used (e.g. keys)
- record persons involved with each step
- record data conversions, modifications, validations...
- record why certain steps were taken



Some more thoughts on the data management plan



data format

Data acquisition

manual

automatic

Data acquisition

manual

automatic

Minimize number of interactions (recordings)

Ideal: double entry & cross check

Data entry interface

Data acquisition

manual

automatic

Minimize number of interactions (recordings)

Ideal: double entry & cross check

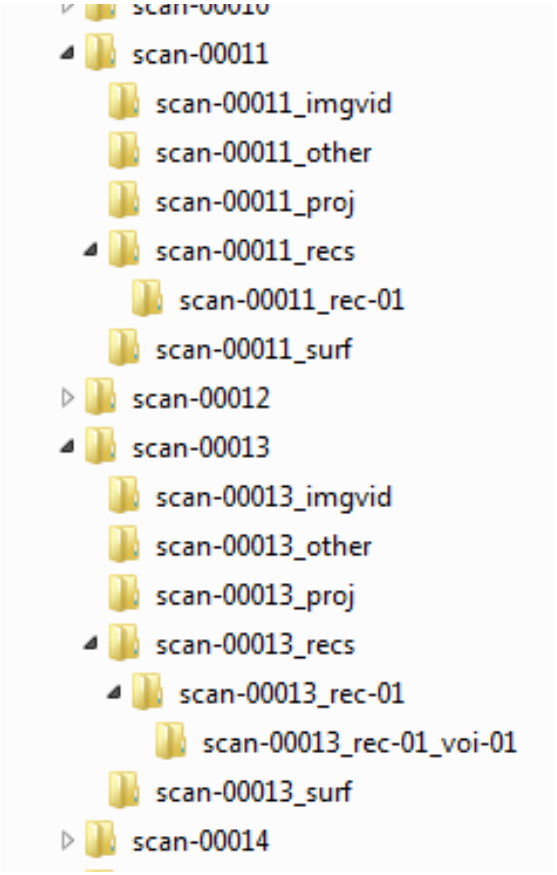
Data entry interface

Recordings correct?

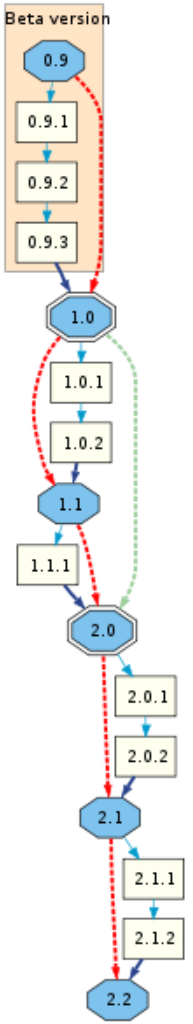
Units included?

Calibration and metadata noted?

File management plan



logical file names
and directory structure



versioning

backup plan

source of versioning image: AzaToth (Wikipedia)

Data management plan created

Consistent codes and labels chosen

Metadata recorded

Data organized logically and consistently

Files named and organized logically

Files in common format (e.g. csv)

Backups made regularly