Reproducible
Research
Workshop

Course in Data Information Literacy

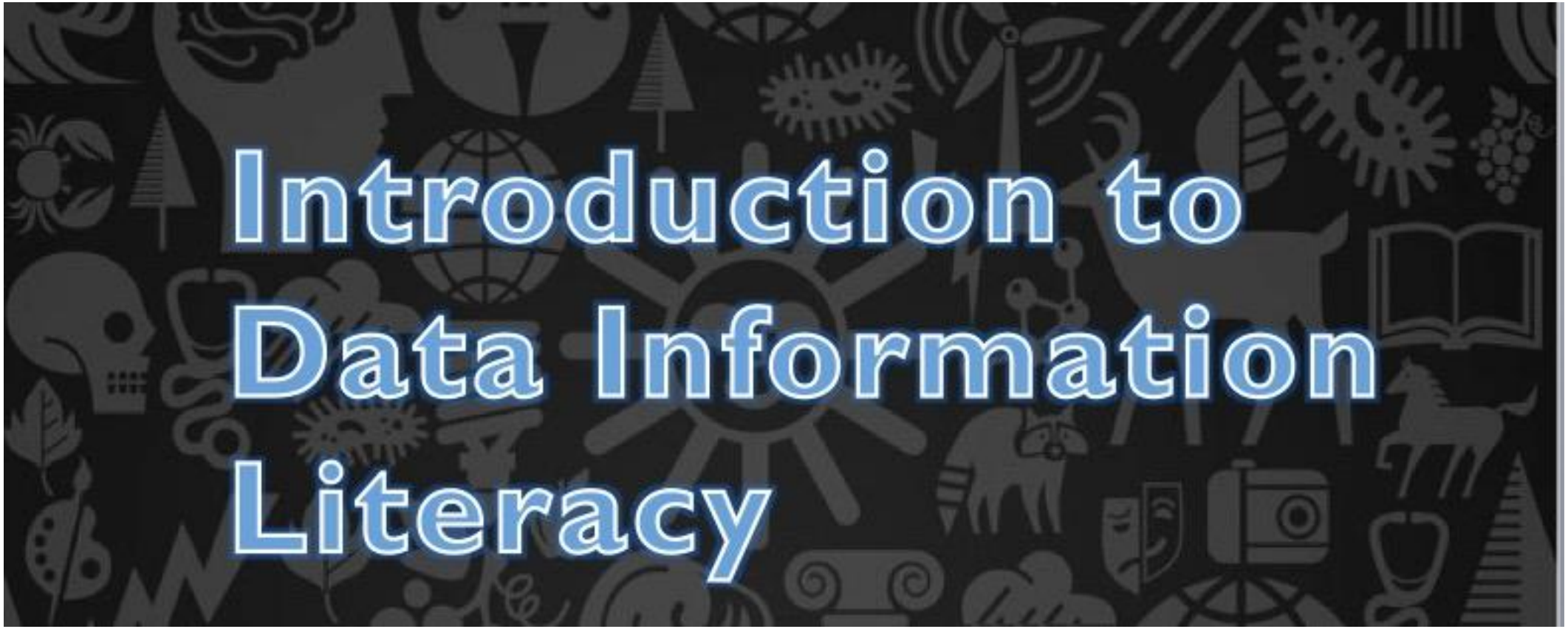
a Progress Report

YOUR NAME: GARY SEITZ

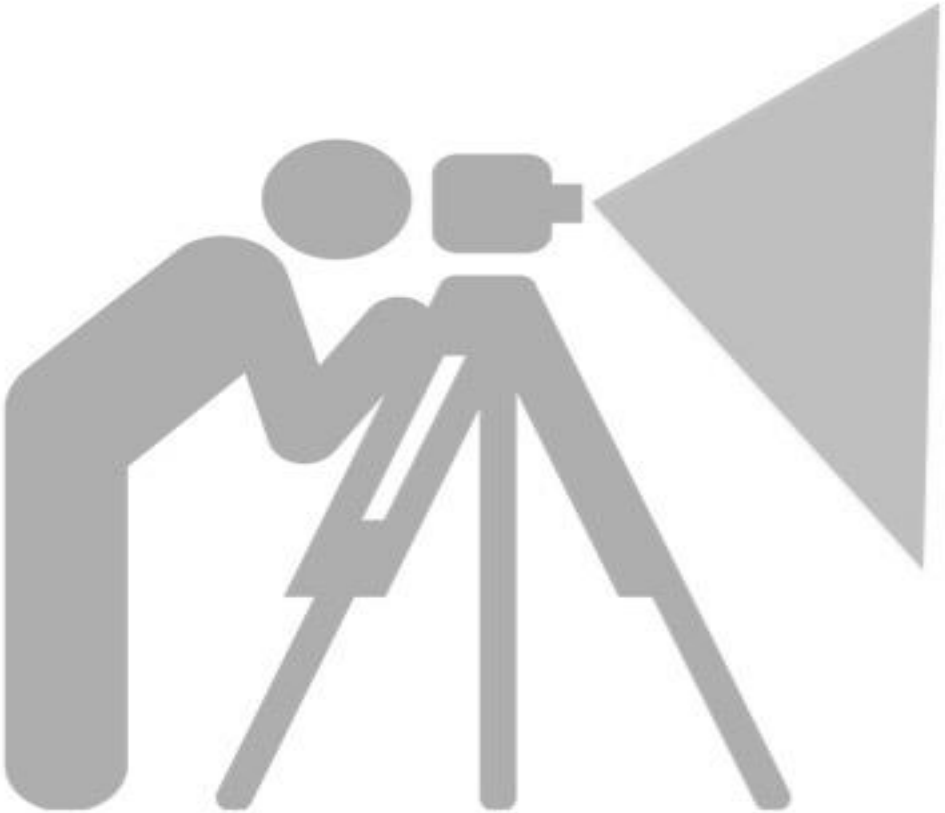
CONTACT: SEITZ@GEO.UZH.CH

Reproducible
Research
Workshop

Lecture 1



Outline



Introductions

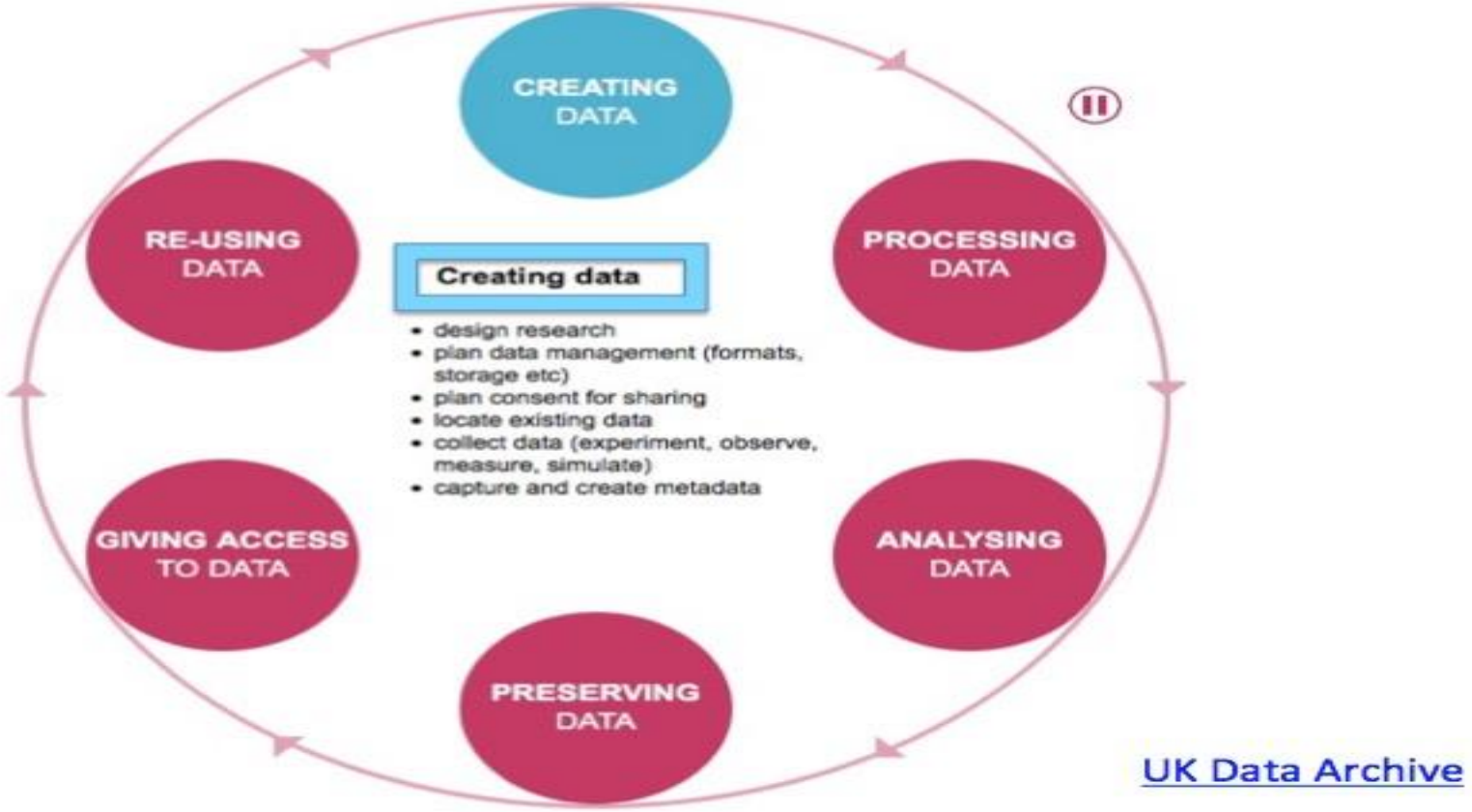
What is/are ‘data’?

The data world around us

The importance of data management

The data lifecycle

Data Lifecycle



UK Data Archive

Summary Lecture 1

The **data deluge** has created a surge of information that needs to be well-managed and made accessible.

The **cost** of not doing data management can be very high.

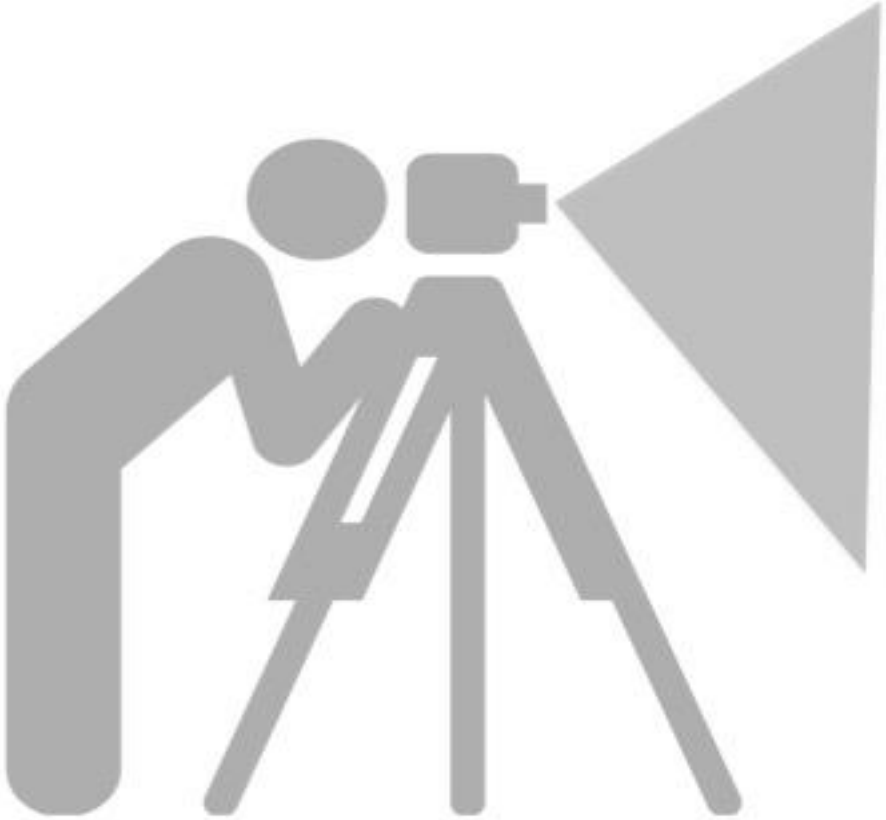
Be cognizant of best practices and tools associated with the data lifecycle to manage your data well.

Many **benefits** are associated with the act of managing data, including the ability to find, access, understand, integrate and re-use data.

Reproducible
Research
Workshop

Lecture 2

Outline



Funding agency context

Data management plan components

Data management examples

Components of a Data Management Plan

1. Information about data & data format
2. Metadata content and format
3. Policies for access & sharing
4. Policies & provisions for reuse
5. Long-term storage and data management
6. Budget

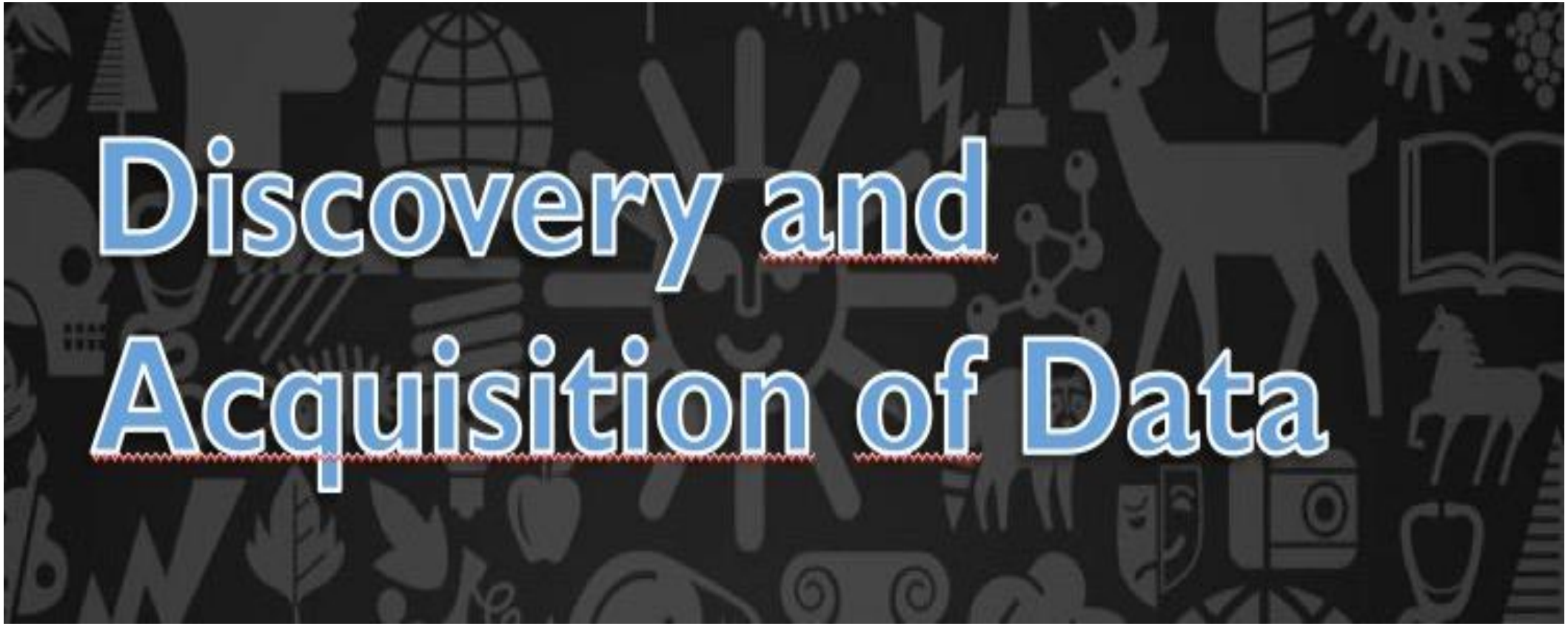
Summary Lecture 2

DMPs are an important part of the data life cycle. They save time and effort in the long run, and ensure that data are relevant and useful for others.

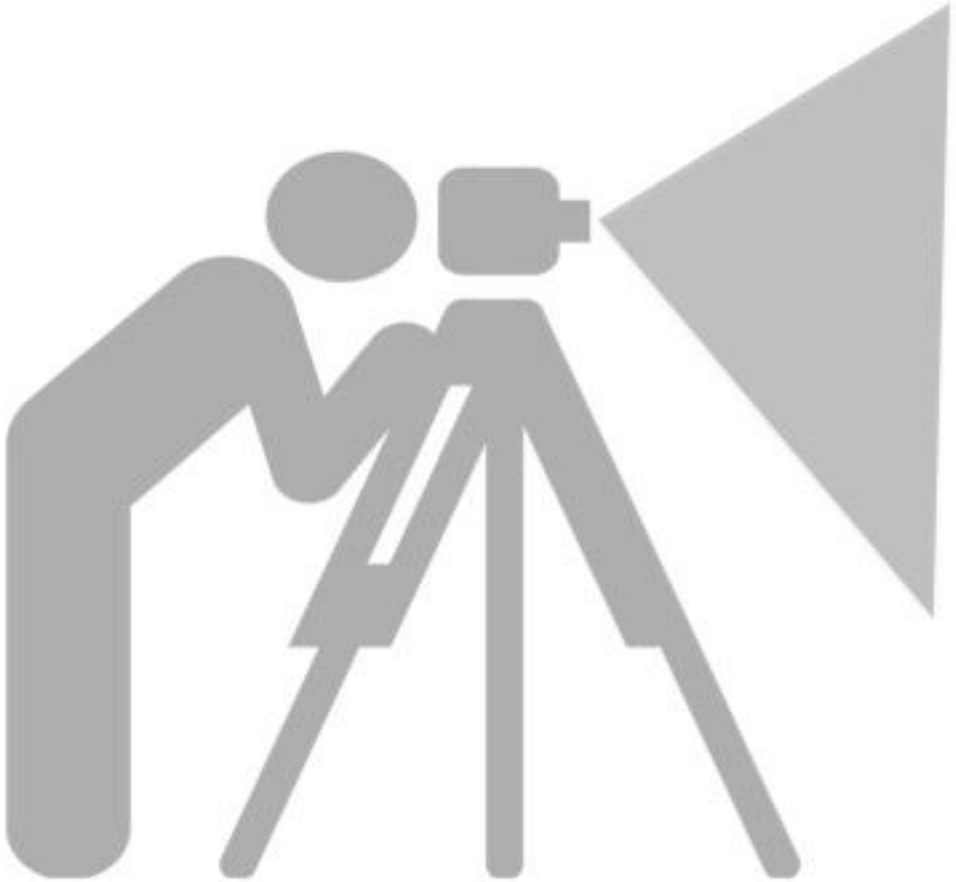
DMPTool
Build your Data Management Plan

Reproducible Research Workshop

Lecture 3



Outline



Data Repositories

Discipline-related repositories

Portals for data publication

Open Data from organizations

Data Repositories

figshare

Dataverse

GitHub

DataCite

Dryad

Protocols.io

re3data.org

Summary Lecture 3

re3data.org is a global registry of research data repositories that covers research data repositories from different academic disciplines

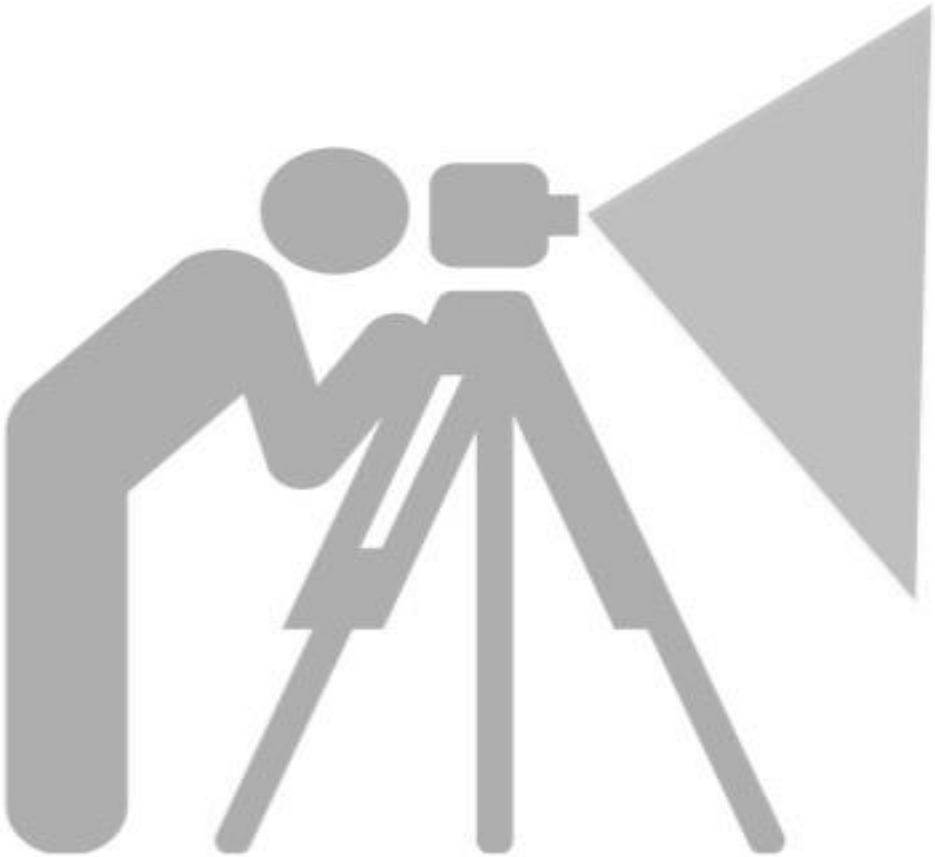
Depending on the research discipline, data can often be accessed in one or more data centers (or repositories) that will provide access to the data

These repositories may have specific requirements

- ✓subject/research domain
- ✓data re-use and access
- ✓file format and data structure, and
- ✓metadata.

Lecture 4

Outline



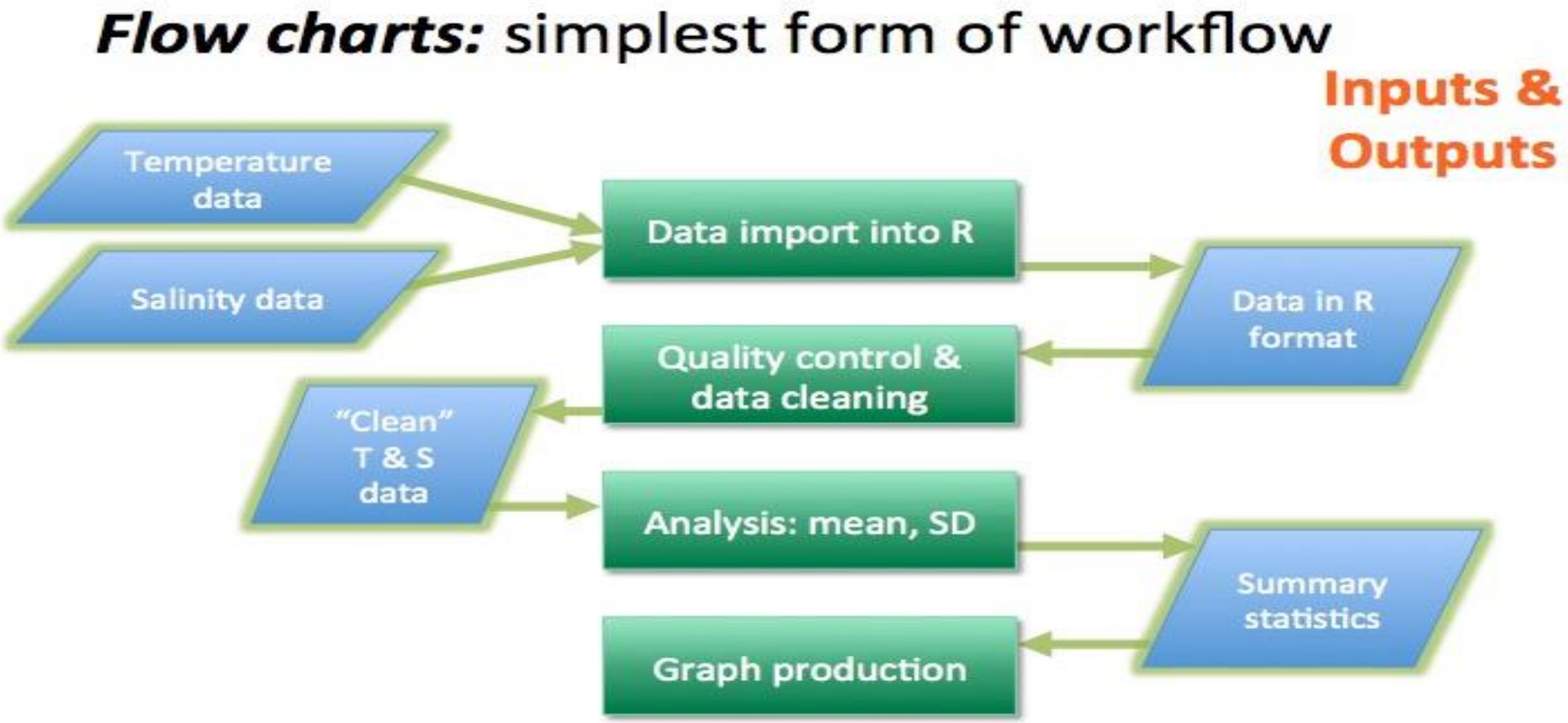
Best practices for creating data & spreadsheet files

Data entry options

Data manipulation options

Analysis and Workflows

Informal Workflows

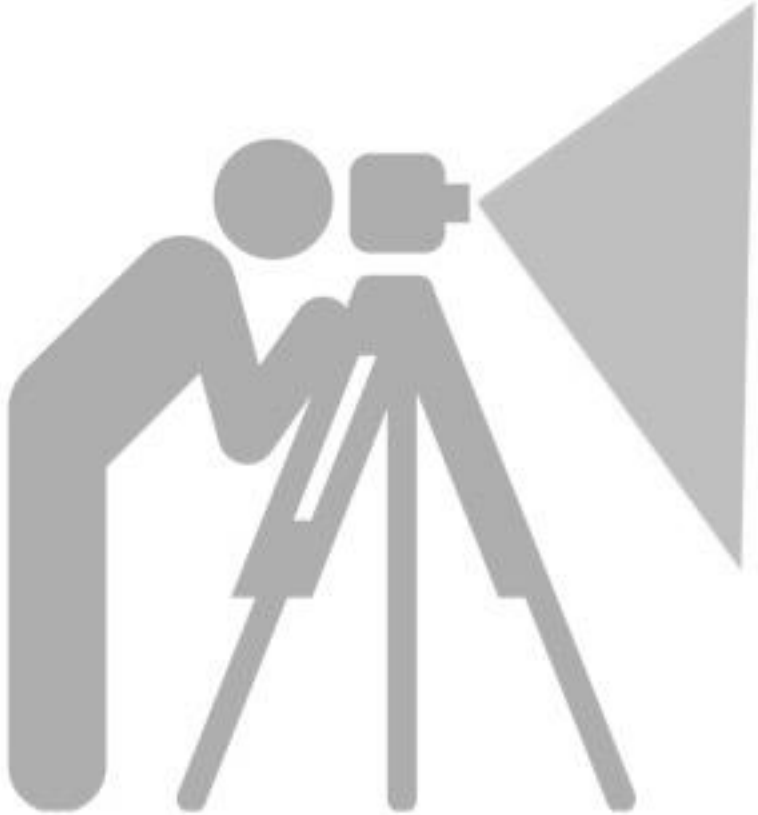


Summary Lecture 4

- ✓ Use of informal or formal workflows for documenting process metadata ensures reproducibility, repeatability, validation
- ✓ Be aware of **best practices** when designing data file structures
- ✓ Choose a data entry method that allows some **validation** of data as it is entered
- ✓ Consider investing time in learning how to use a database if datasets are large or complex

Lecture 5

Outline



File-naming conventions

Data organization

Documenting your process

Keeping a [lab] notebook

File naming strategies

Order by date:

19550412_notes_MassObs.docx
19550412_questionnaire_MassObs.pdf
19631215_notes_Gorer.docx
19631215_questionnaire_Gorer.pdf

Order by type:

Notes_Gorer_19631215.docx
Notes_MassObs_19550412.docx
Questionnaire_Gorer_19631215.pdf
Questionnaire_MassObs_19550412.pdf

Order by subject:

Gorer_notes_19631215.docx
Gorer_questionnaire_19631215.pdf
MassObs_notes_19550412.docx
MassObs_questionnaire_19550412.pdf

Forced order with numbering:

01_MassObs_questionnaire_19550412.pdf
02_MassObs_notes_19550412.docx
03_Gorer_questionnaire_19631215.pdf
04_Gorer_notes_19631215.docx

Summary Lecture 5

When naming & organizing

your files and folders…

be thoughtful

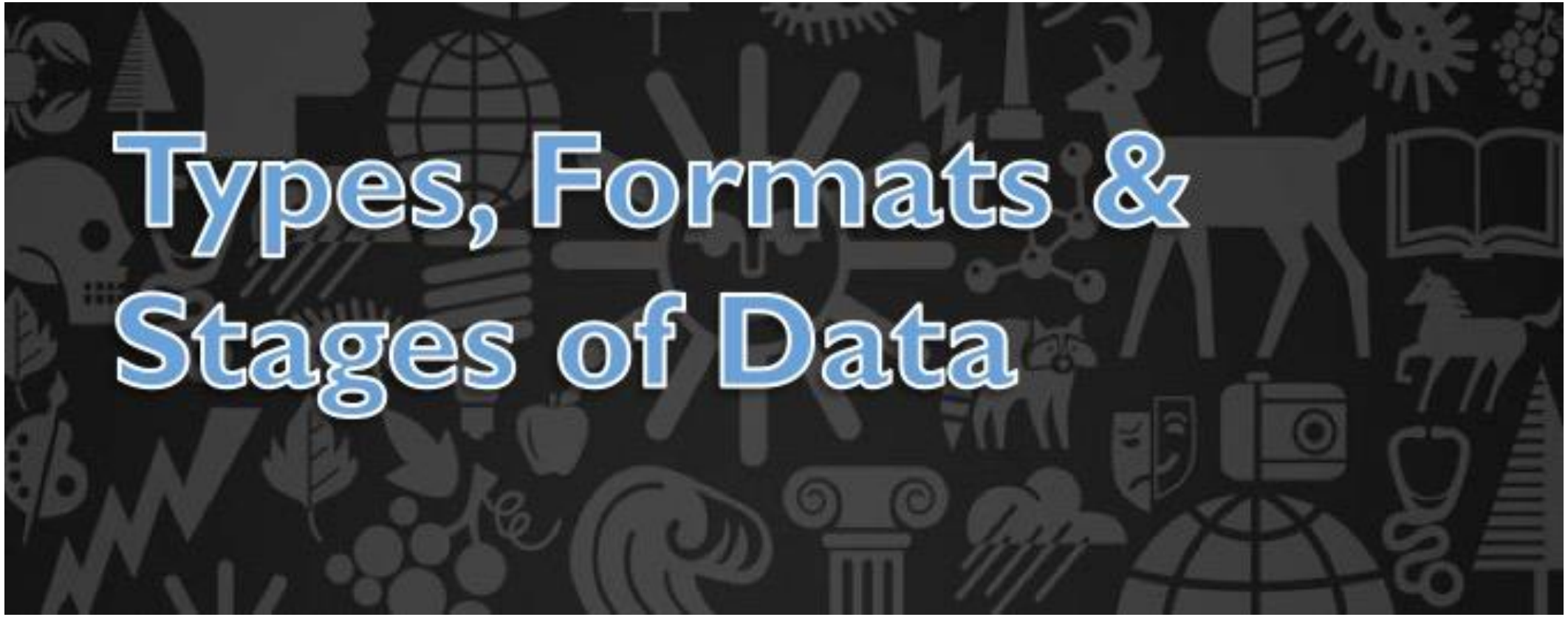
be consistent

document your approach

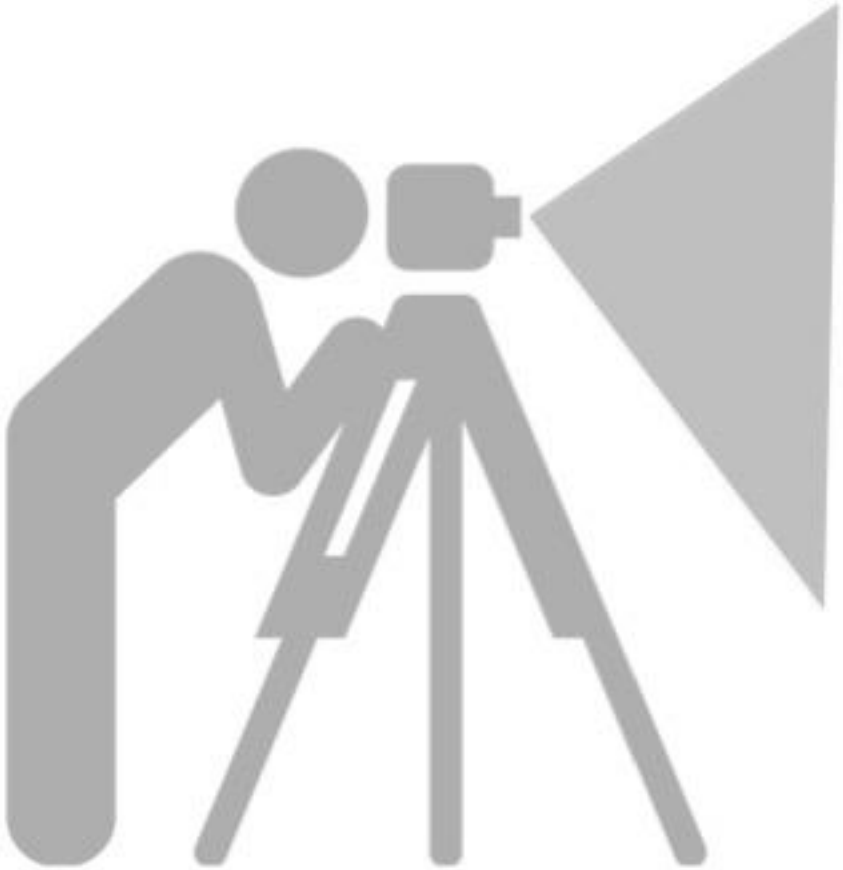
Write down

All The Things

Lecture 6



Outline



Data types & formats

Research lifecycle & stages of data

Preferred Formats

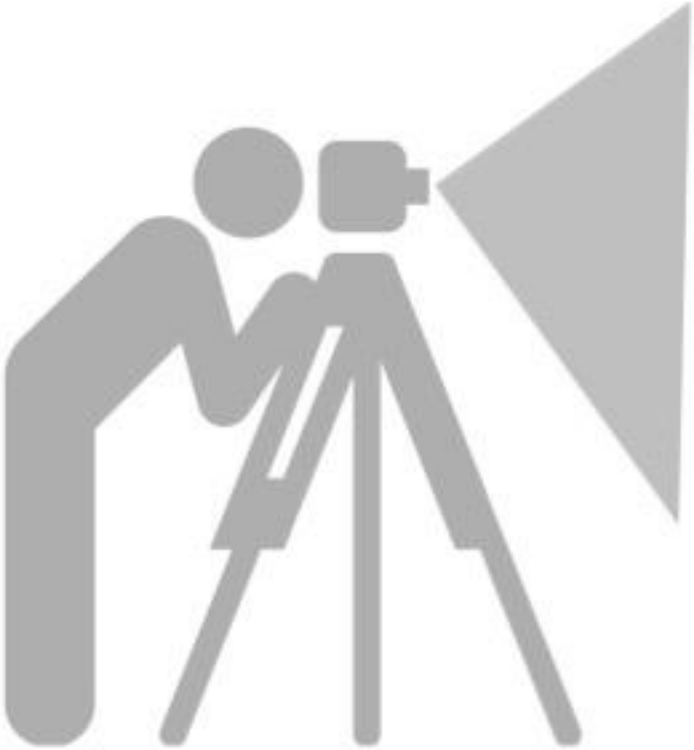
File type	Suitable for use over more than ten years	Use limited to ten years	Not suitable for archiving
Raster image (bitmap)	– TIFF (*.tif) (uncompressed, preferentially TIFF 6.0, Part 1: baseline TIFF) – PNG (uncompressed) – JPEG2000 (lossless compression)	– TIFF (*.tif) (compressed) – GIF (*.gif) – BMP (*.bmp) – JPEG/JFIF (*.jpg) – JPEG2000 (lossy compression) (*.jp2)	
Vector graphics	– SVG without JavaScript binding (*.svg)		– Graphics InDesign (.indd), Illustrator (.ait) – Encapsulated Postscript (EPS)
Audio	– WAV (*.wav) (uncompressed, pulse-code modulated)	– Advanced Audio Coding (*.mp4) – MP3 (*.mp3)	
Video	– Motion JPEG 2000 (ISO/IEC15444-4) (*.mj2) – AVI (uncompressed, motion JPEG) (*.avi) – QuickTime Movie (uncompressed, motion JPEG) (*.mov)	– MPEG-1, MPEG-2 (*.mpg,*.mpeg, wrapped into the container format AVI or MOV) – MPEG-4 (H.263, H.264) (*.mp4, wrapped into the container format AVI or MOV)	– Windows Media Video (*.wmv)

Summary Lecture 6

- ✓ Programs and file formats change over time such that old files may become difficult to read.
- ✓ Files in rare formats should be converted into common formats whenever possible.
- ✓ Files should not be password protected, encrypted or compressed
- ✓ File formats should be very common and, if possible, follow standards that are open and not proprietary
- ✓ For storage over more than ten years, we recommend the file formats PDF/A, ASCII text, TIFF, PNG, SVG and JPEG2000
- ✓ For large data collections you can get an overview of your file formats using the free JAVA application DROID

Lecture 7

Outline



Definition of metadata

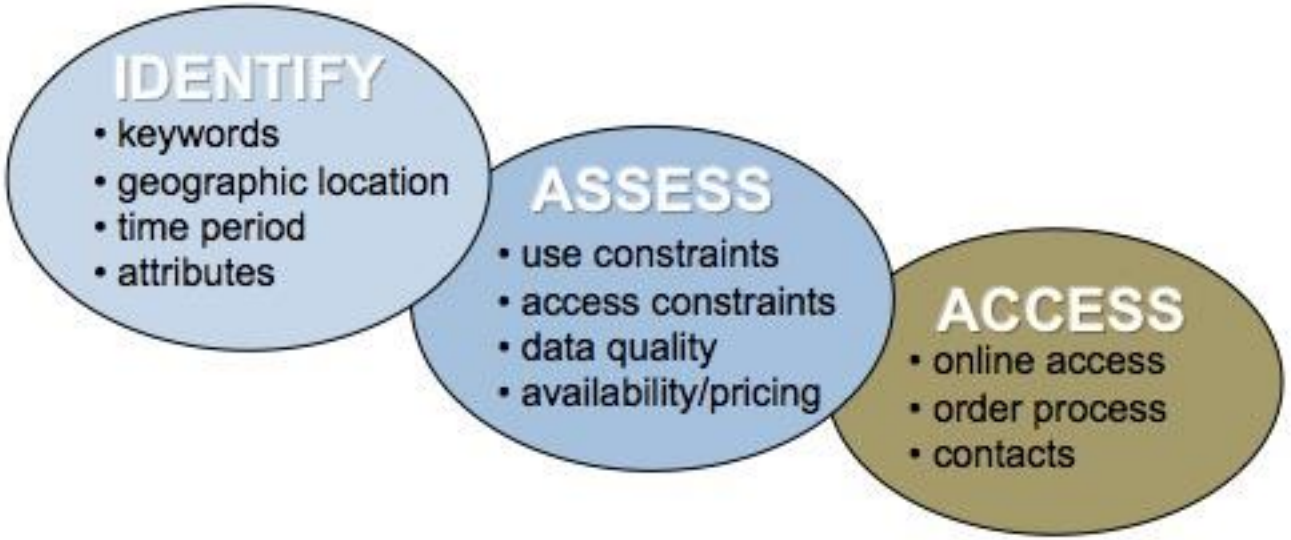
Examples of metadata standards

Illustrate the value of metadata to data users

Describe the utility of metadata

Distribution: data discovery

The descriptive content of the metadata file can be used to identify, assess, and access available data resources.



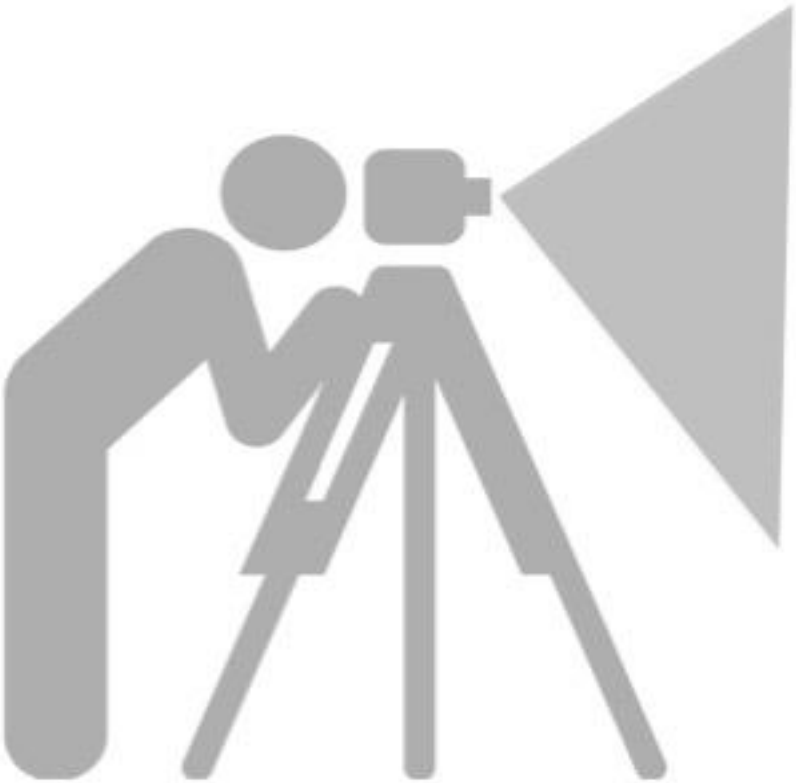
Summary Lecture 7

- ✓Metadata is documentation of data
- ✓A metadata record captures critical information about the content of a dataset
- ✓Metadata allows data to be discovered, accessed, and re-used
- ✓A metadata standard provides structure and consistency to data documentation
- ✓Standards and tools vary – select according to defined criteria such as data type, organizational guidance, and available resources
- ✓Metadata is of critical importance to data developers, data users, and organizations
- ✓Metadata can be effectively used for:
  - ✓ data distribution
  - ✓ data management
  - ✓ project management
- ✓Metadata completes a dataset.

Creating robust metadata is in your OWN best interest!

Lecture 8

Outline



Why?

Where to store data

Data backup

Data security

Data Preservation

Backups vs. Archiving

Reproducible Research Workshop

Backups

- Used to take periodic snapshots of data in case the current version is destroyed or lost
- Backups are copies of files stored for short or near-long-term
- Often performed on a somewhat frequent schedule

Archiving

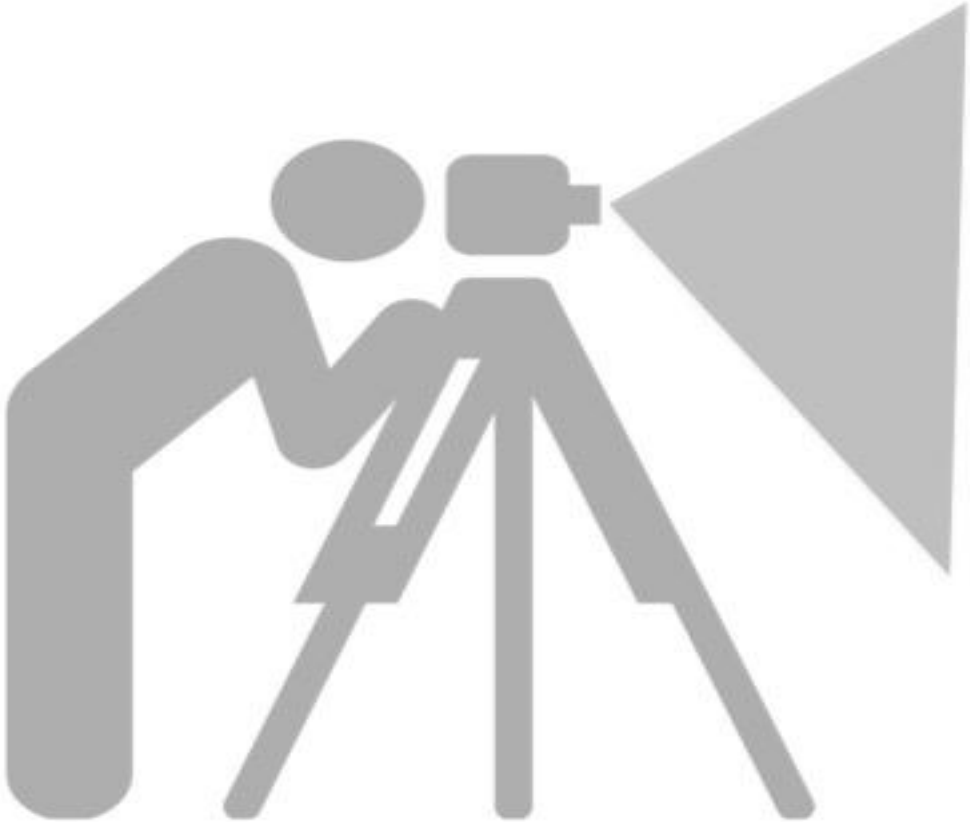
- Used to preserve data for historical reference or potentially during disasters
- Archives are usually the final version, stored for long-term, and generally not copied over
- Often performed at the end of a project or during major milestones

Summary Lecture 8

- ✓Backups refer to creating copies of original files while archives involve the preservation of files
- ✓There are many reasons we need to perform backups but primarily to prevent data loss
- ✓One needs to consider how often to perform backups, where to backup, and accessibility to backups when you need them and how long to keep the files
- ✓Check for backups on outdated media and test backups often!
- ✓Data preservation more than just backing up and archiving your files
- ✓Evaluate and refresh storage regularly
- ✓Protect the integrity of your data at the file level
- ✓Protect the hardware and software systems you use

Lecture 9

Outline



Basic archival processes

Need for conversion to standard formats

Options for a long-term sustainable preservation

Costs & timelines for data storage

Select archive location

Considerations

- Costs
- Size of dataset
- Public vs. private access
- Length of preservation
- Hands-on vs. hands-off
- Security of platform

Locations

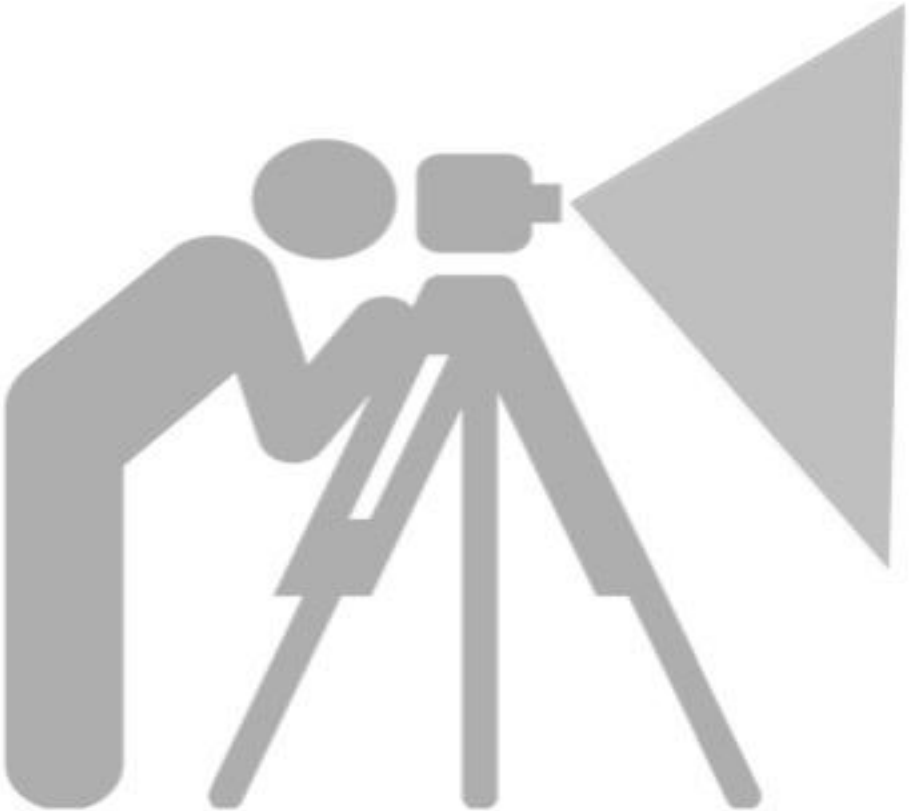
- Individual
- Department/College
- University-wide
- Discipline-specific
- 3rd-party
  - Archive vs. sharing mechanism

Summary Lecture 9

- ✓ Data preservation has many potential benefits:
  - ✓ Enable longitudinal and synthesis studies
  - ✓ Leverage investments in data collection
- ✓ Additional considerations
  - ✓ Preservation of data in multiple forms - i.e. raw, processed, derived, etc - may be warranted in many circumstances.
    - ✓ Which version(s) to keep?
    - ✓ How to make relationships among versions clear?
  - ✓ Considerations of cost and reproducibility are key in considering policies for preservation of experimental data.
    - ✓ How to assess the long-term value of data?
    - ✓ What documentation is necessary to enable data replication?

Lecture 10

Outline



Benefits of sharing data

Issues/obstacles related to reuse and sharing of data

Understand open access

Understand data reuse policies from funding agencies

Data citation

Value of Data Sharing to the Public

A better informed public yields better decision making with regard to:

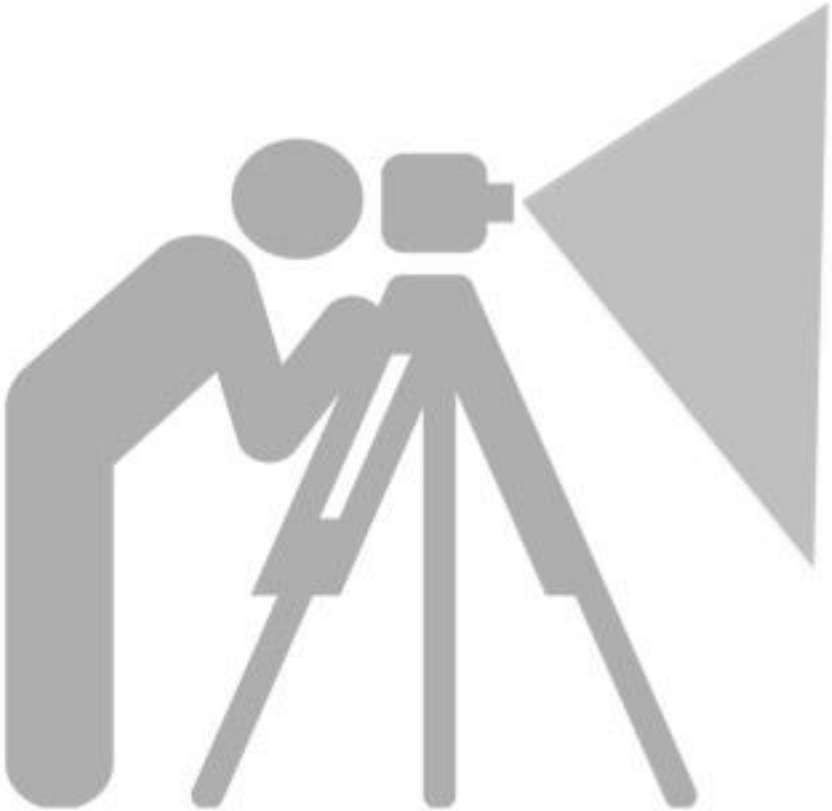
- Environmental and economic planning
- Federal, state, and local policies
- social choices such as use of tax dollars and education options
- personal lifestyle and health such as nutrition and recreation

Summary Lecture 10

- ✓ Data sharing adds value to the data
- ✓ It is the responsibility of the researcher to share their data
- ✓ Metadata supports data accountability, liability, and usability
- ✓ Sponsors expect, some require, data to be shared
- ✓ Data sharing is essential to the advancement of science
- ✓ Data Citation makes it easy for others to attribute your data directly to you

Lecture 11

Outline



Identify ethical, legal, and policy issues

Define copyrights, licenses and waivers

Understand reasons behind data restrictions

Discuss ethical considerations

Deidentification of Research Data

The process of anonymizing data to protect the identity of the participants and to remove other private information.

General Guidelines:

- Mark replacements of text clearly, either by using [brackets] or tags: <anon> … </anon>
- Keep a secure copy of the non-anonymized data.
- Create a log of all the replacements, aggregations, or removals made in each data file. **Store this log file separately from the de-identified data.**

Summary Lecture 11

- ✓ Know who can claim ownership over products
- ✓ Assign licenses or waivers appropriately
- ✓ Behave ethically and in accordance with established community norms
- ✓ Respect the licenses or waivers assigned
- ✓ Protect privacy and confidentiality
- ✓ Know what restrictions and liabilities apply to products and processes

Thank you for all your comments!