PUBLISHING AND ARCHIVING DATA

- **Mikko Mäkelä** **DataSupport**

CONTENTS

MOTIVATION AND TERMINOLOGY

Research data is always valuable: Money, time and work has been invested in it

To get full benefits from it, valuable data should eventually be made available for all

Open research data benefits both researchers and scientific community

Data publication / opening is the process of making data generated from research available to all

Goal is to promote reuse of the data, timespan is around 5-15 years

Data long-term preservation / archiving is the long-term storage of such data and methods. The data will often need to be published as well.

Goal is the preservation and integrity of the data

Both require careful preparation, planning and varying amounts of additional effort

FATE OF THE DATA AFTER THE STUDY

What happens to your data after the papers have been published?

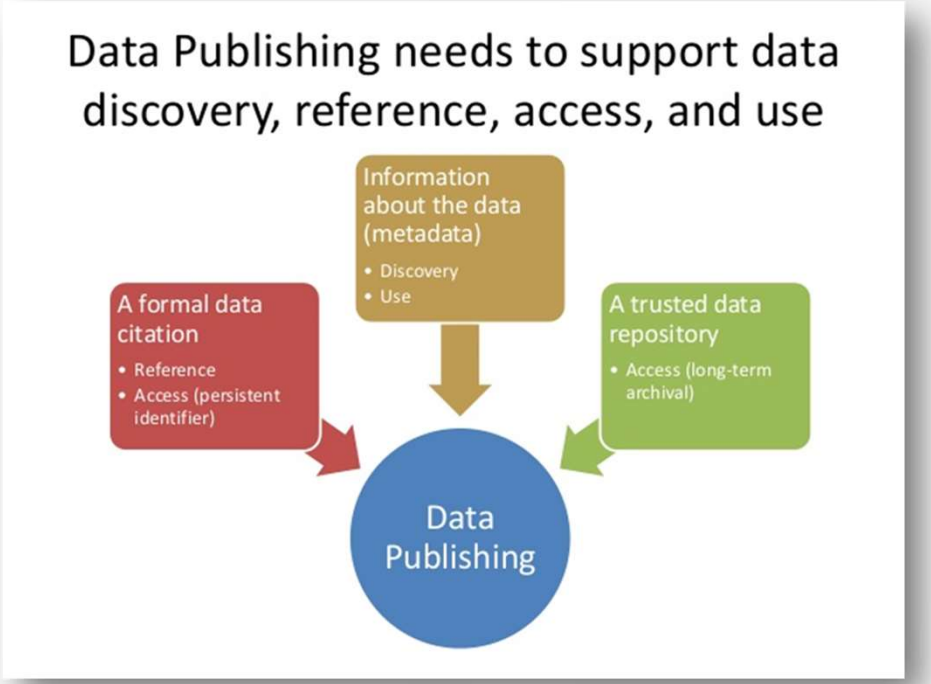
When planning your research, you must also plan what happens to your data <u>after</u> the research!

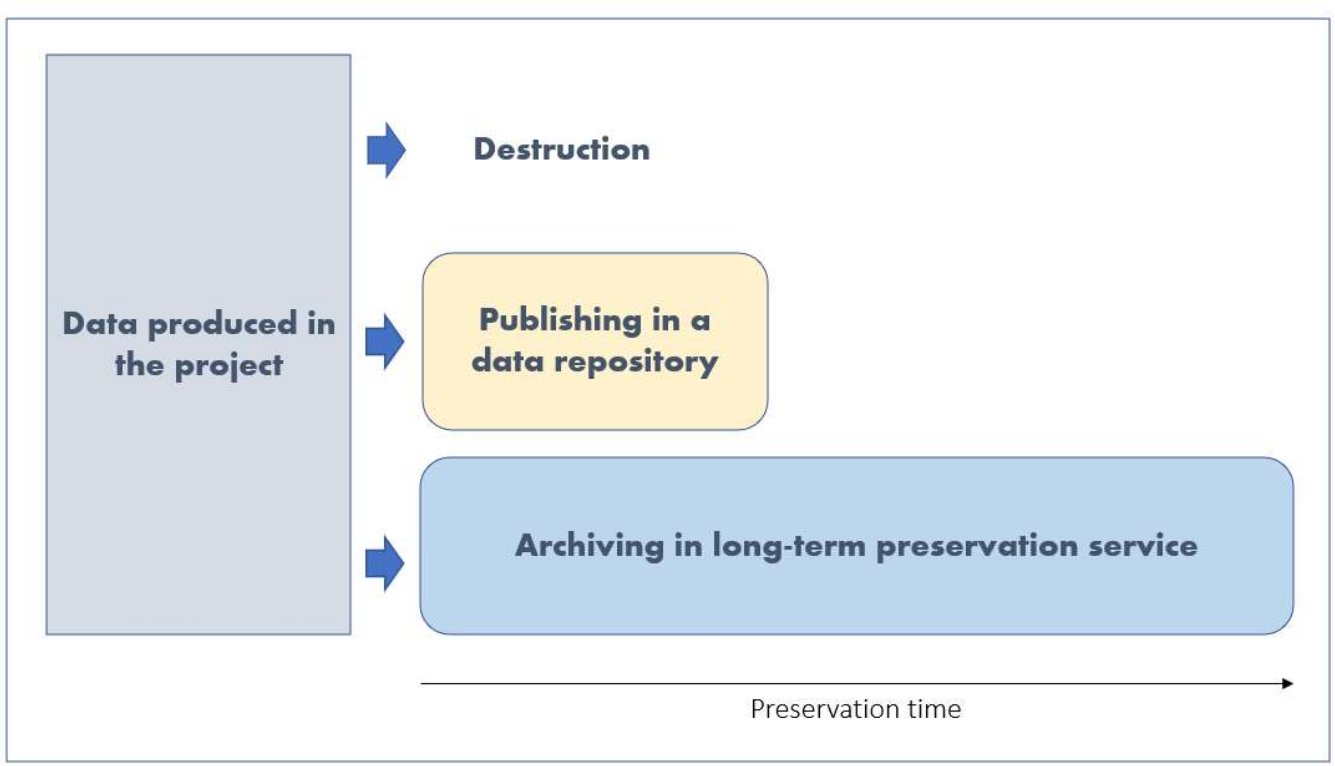
- Make agreements and contracts accordingly
- Inform the study subjects
- Plan your metadata collecting

Sometimes, research data cannot be published as a whole

At least the metadata can be opened

”As open as possible, as closed as necessary”




Data produced in the project
Destruction
Publishing in a data repository
Archiving in long-term preservation service
Preservation time

1. DATA REPOSITORIES – FOR PUBLISHING AND REUSE

- Large database infrastructure that collects, manages, and stores datasets for data analysis, sharing and reporting
- Repositories can host actual datasets or they can be specialised for metadata
- Preservation time varies: 5 - 15 years

Benefits of data repositories

Data is preverved beyond work contract length;
Data is accessible from everywhere
Easy to discover by search engines;
Citation system, licenses and persistent identifiers
Increase visibility of your work;
Funders may require data to made available online

HOW TO CHOOSE A REPOSITORY – SOME IDEAS

- Prefer curated repositories. The data curators can help you to secure the quality of metadata, choosing file formats etc.
- Prefer a disciplinary repository or repositories which include similar datasets to your own! Your data is likely to be found better when you are in good company
- Find out what are the costs of the repository (if any)! Does your research budget cover for the costs?
- What is the physical storage location of data (EU, GB or USA)? There may be issues with agreements, funder's requirements etc.
- Do the repository provide the dataset a PID? DOI, URN, ARK or handle promote citability
- What is the default license? Some repositories require using CC0 (giving up the copyright)
- Is long-term preservation guaranteed or not? Preservation length is stated in the service agreement
- OpenAire guide: How to find a trustworthy repository for your data?

RE3DATA - REGISTRY OF RESEARCH DATA REPOSITORIES

- Zenodo and Figshare are general repositories for all fields of science – it might be better to publish in a repository which is dedicated to certain fields of science and with similar data
- re3data.org is a global registry of research data repositories from different academic disciplines
- It presents repositories for the storage and access of datasets to researchers, funding bodies, publishers and scholarly institutions.
- A tool for the easy identification of appropriate data repositories to store research data.
- https://www.re3data.org/

LICENSING YOUR WORK

License informs the users how your work can be reused or shared

Licensing makes wider distribution possible

For research data UH recommends license which is open as possible

The Creative Commons -licenses (CC) have become a standard for open publishing

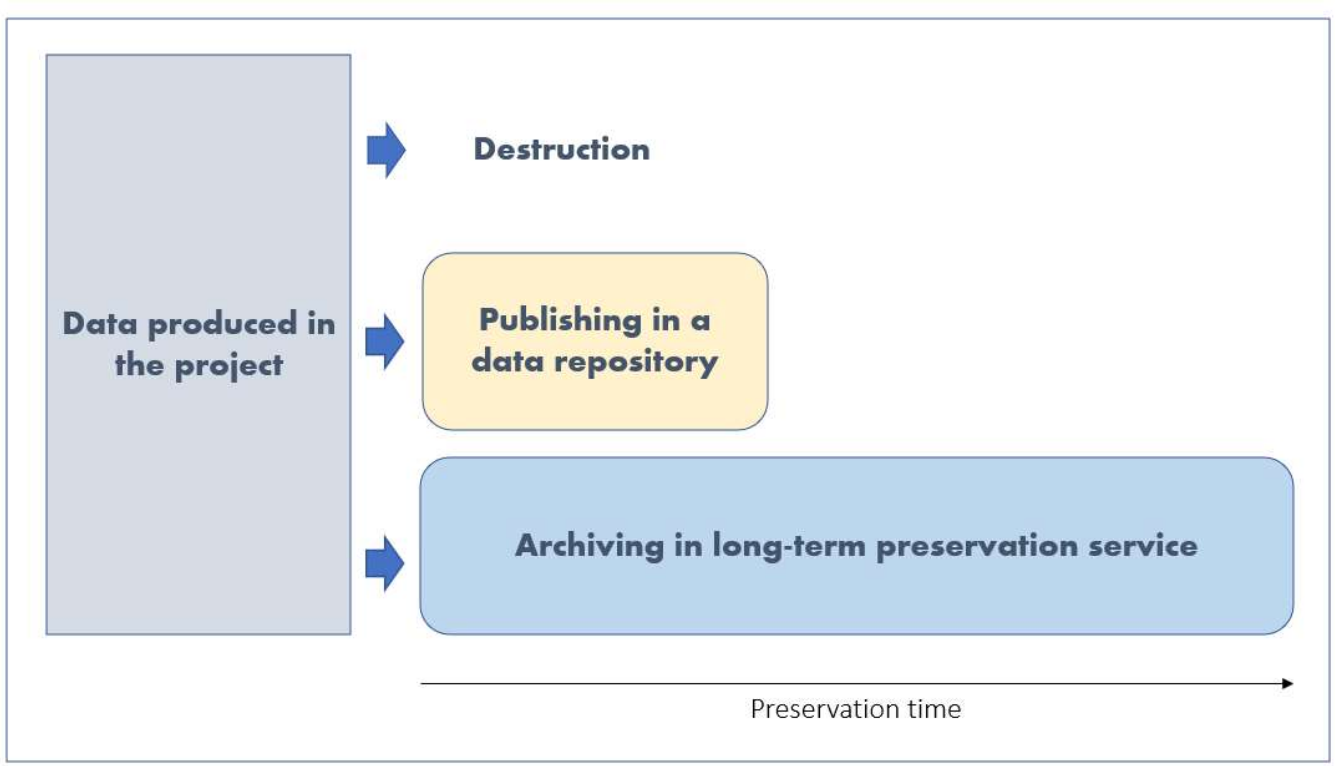
Sotfware use different licensing (MIT license and GNU General Public License)

You can choose the license when you publish your work, but not in commercial publishing platforms like Research Gate

On different CC-licenses, see our License guide:
http://libraryguides.helsinki.fi/oa/eng/license

DATA JOURNALS

- Publishing dataset in a peer-reviewed data journal can increase the visibility and usability for the data
- Usually the dataset itself must be uploaded to a repostery
- There are general and disciplinary data journals
- How is the findability of the data? Does the supplements show in search engines?
- Examples of generic data journals: Scientific Data; Data in Brief; Data Science Journal


Data produced in the project
Destruction
Publishing in a data repository
Archiving in long-term preservation service
Preservation time

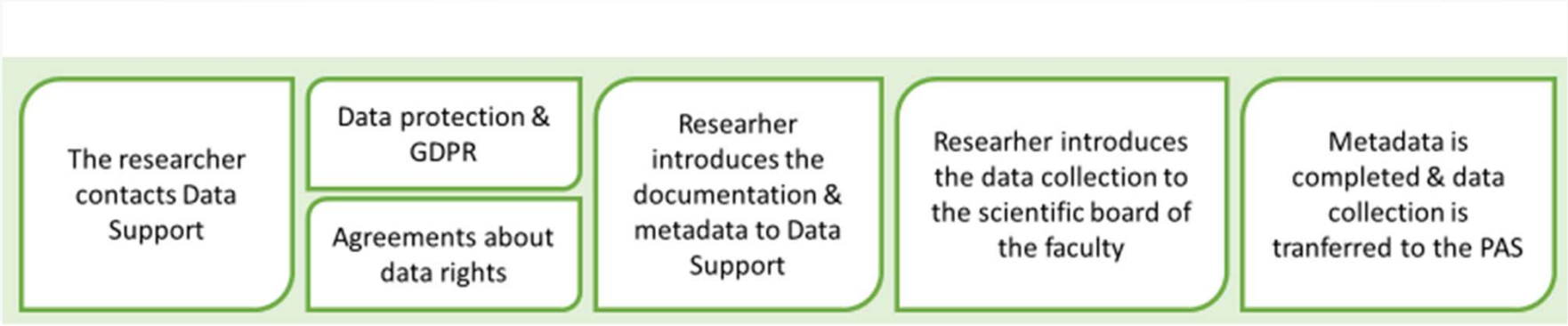
2. ARCHIVING THE DATA – LONG-TERM PRESERVATION

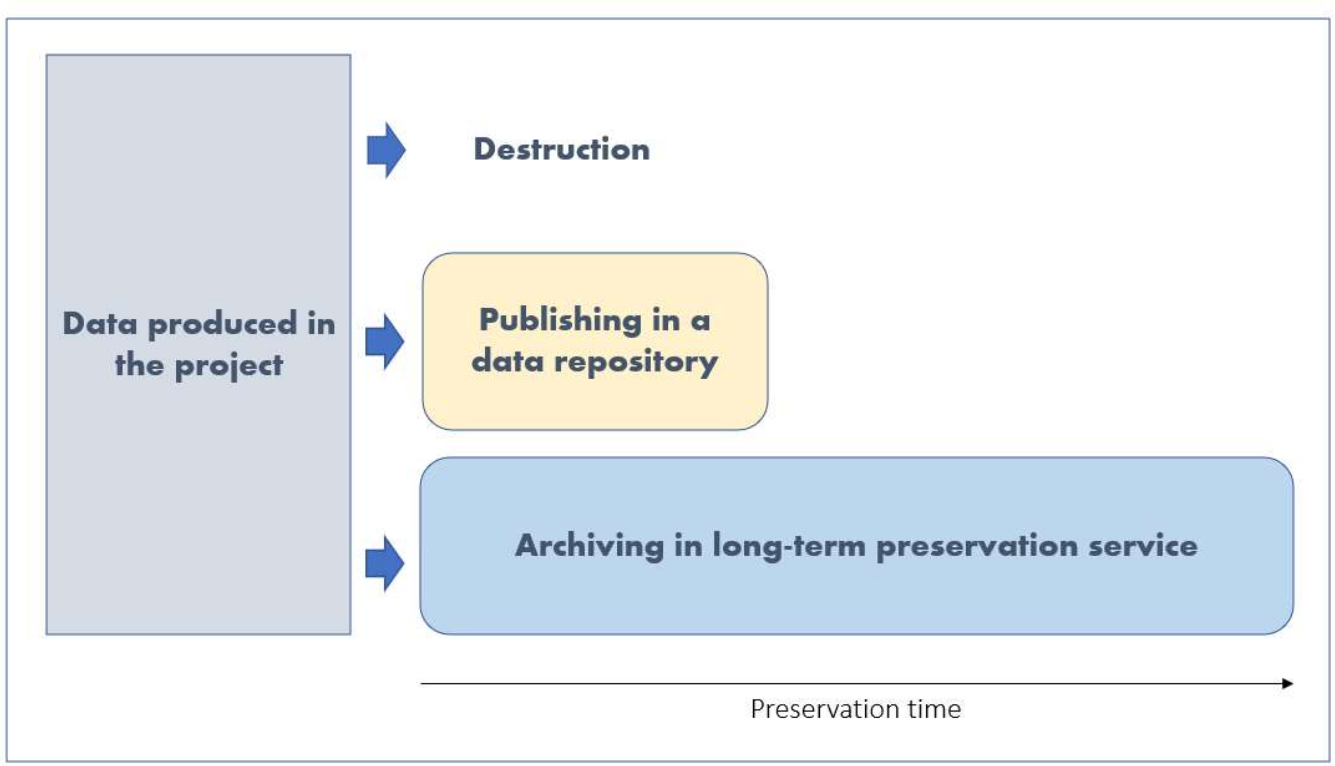
- Long-term preservation over tens or hundreds of years (over generations)
- The goal of archiving is preservation, not so much reusability
  - Distinct from publishing – archived data should be published as well
- Challenges: aging of file formats and hardware, legislation is complex
- Long-term preservation is an active process which requires special resources and expertise
- Advanced requirements for metadata = datasets should be able to explain themselves

LONG-TERM PRESERVATION IN UH

- Curated service for datasets with national importance
- Administration tasks by UH, tech by CSC
- Requires co-operation from legal and IT support as well as the library
- Desicion for preservation is made by the scientific board of each faculty

Long-term preservation service




Data produced in the project
Destruction
Publishing in a data repository
Archiving in long-term preservation service
Preservation time

3. DATA DESTRUCTION

As a default research datasets should be published or archived

However, this is sometimes impossible and data will have to be destroyed in controlled manner

- Sensitive data
- Personal data when anonymisation is not possible
- Raw data from video or audio recordings (if not necessary)
- Study permissions do not allow preservation
- Informing the study subjects determines preservation time

Technically straightforward procedure done manually

- Make sure you delete all copies
- Take into account the backups in UH storage systems

TAKEHOME MESSAGE

Start thinking about data publishing in advance

- Plan metadata collection and study permissions accordingly

Decide the appropiate preservation times for your data

- Preservation for reuse -> online repository
- Preservation for over generations -> long-term preservation service
- Destruction at the end of the project