Publishing Research Data

Online workshop for researchers

28 January 2022

Dragan Mišić,
University of Niš, Faculty of Mechanical Engineering

Goals

1. **To understand why it is necessary to publish data**
2. **To learn how and where to publish data**
3. **To learn how to cite data**

euraxess
RESEARCHERS IN MOTION

Content

- What is open research data?
- What are the benefits of opening data?
- What metadata is?
- FAIR guiding principles
- Data Management Plan
- Where to publish data?
- Hot to publish data?
- How to cite data?
- Exercise: Search for open data

euraxess RESEARCHERS IN MOTION

What is open research data?

- **Data that can be freely accessed, reused, remixed and redistributed, for purposes of academic research and teaching and beyond**
- Ideally, open data have **no restrictions on reuse or redistribution**, and are **appropriately licensed** as such
- At most, the requirement to **attribute** and **share alike** are present

euraxess
RESEARCHERS IN MOTION

Benefits of open data

Researchers

- Greater discoverability
- Increased efficiency
- Attracts funding and support
- New collaborations

Funders

- Increased visibility and reuse of funded research
- Greater funding impact
- Greater ROI

General Public

- Self-improvement
- Increased transparency
- Greater engagement in science and research

Organisations/NGOs

- Enhanced access to research
- Better information sharing
- More efficient advocacy

National governments

- Data-driven decision making
- Reduced government cost
- More effective and efficient services

euraxess RESEARCHERS IN MOTION

Why publish data (benefits for researchers)

- Allowing the replication of experiments
- Potential to introduce new collaborations
- Boosting your reputation within community
- Increasing visibility of the research
- Increase citations
- **as open as possible and as closed as necessary**.

Concerns about open data

- **Confidentiality** issues
- I**ntellectual property** concerns
- **Losing full control** of the data
- Using data for similar research when original author's paper **is not yet published**
- **Digging for errors** in data analysis
- Possible **solutions**
  - Managing permission settings for reuse
  - Avoiding misinterpretation by showing transparency

Metadata

- Metadata is structured information that makes it easier to **retrieve, use or manage** an information resource.
- Metadata describes a dataset and its structure, and helps users discover it.
- Typical metadata:
  - title,
  - who published the dataset,
  - when it was published,
  - how often it is updated and what license is associated with the dataset.
- Descriptive metadata vs structural metadata

euraxess RESEARCHERS IN MOTION

Metadata types

- Metadata that provides an overview of the data.
  - helps people find the data through internet searches, while navigating your portal, or even while navigating other data portals which might include your catalog.
- Metadata that provides details about specific parts of your data.
  - enables people to use your data effectively, by helping them understand the various elements it includes and potential limitations

Dataset metadata

- **Title** (or Name): Human-readable name for the data. It should be in plain English and include sufficient detail to facilitate search and discovery. Acronyms should be avoided.
- **Description**: Human-readable description (e.g., an abstract) with sufficient detail to enable a user to quickly understand whether the asset is of interest.
- **Category** (or Theme): Main thematic category of the dataset, usually chosen from a predefined list. Some open data portals limit a dataset to one category; others allow multiple.
- **Keywords** (or Tags): Tags (or keywords) are generally single words which help visitors discover the data; please include terms that would be used by technical and non-technical users. Keywords can also be used by recommendation engines to help visitors discover similar datasets.
- **Modification Date**: The most recent date on which the dataset was changed, updated, or modified.
- **Contact Information**: The name and email address of the publisher of a dataset.
- **License**

Column metadata

- provide important details about the data which the column contains.
- **Name**: Human-readable name of the column.
- **Description**: Human-readable description of the column's contents.
  - should include how values in this column are created or updated; address any data quality concerns, such as unexpected or unusual values;, and explain any meanings which might be stored as codes,often used for record classification, and more frequent in source data systems designed for limited storage space.
- **Data Type**: helps improve the consistency and quality of data (text, numbers, dates/times, booleans (yes/no or true/false))
- **Required**: Specifying whether a value is required in the column for every row in the table helps improve the quality of data

Dataset metadata example

- **Authors**: Misic Dragan, Miroslav Trajanovic
- **Date of publication**: 12 May 2019
- **Title**: Time required for typing numbers
- **Description**: The goal of this test is to see how much time it is needed for one keystroke of a number on the keyboard. …
- **Keywords**: typing numbers, human-computer interface
- **License**: Creative Common Attribution 4.0
- **Contact information**: misicdr@gmail.com

Column metadata example

Name	Description	Data type	Required
Errors per ten digits	Number of errors per 10 entered digits	Number	Yes
Time for one digit	Time needed for entering one digit	Number	Yes
User ID	User identification	Number	Yes
Test date	Date and time of thest	Date	Yes
Gender	User gender. This field can have two values: male and female	String	Yes
Right hand	This field can have values 0 and 1. The value 0 indicates that the user's dominant hand is right. 0 indicates that dominant hand is left.	Number	Yes
Birth year	Year of birth of user	Number	Yes

FAIR data

- **Findable:** easy to find the data and the metadata for both humans and computers. Enabled by machine-readable persistent identifiers (PIDs) and metadata
- **Accessible:** data can be retrieved using open protocols, possibly including authentication and authorization
- **Interoperable:** can be combined and used with other data or tools
- **Re-usable:** well-described so that they can be replicated and/or combined in different settings

To be Findable

- (meta)data are assigned a **globally unique and persistent identifier**
- data are described with **rich metadata**
- metadata clearly and explicitly **include the identifier** of the data it describes
- (meta)data are **registered or indexed** in a searchable resource

To be Accessible

- (Meta)data are retrievable by their identifier using a **standardised communications protocol**
  - The protocol is open, free, and universally implementable
  - The protocol allows for an authentication and authorisation procedure, where necessary
- **Metadata are accessible**, even when the data are no longer available

To be Interoperable

- (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- (Meta)data use vocabularies that follow FAIR principles
- (Meta)data include qualified references to other (meta)data

To be Resuable

- (Meta)data are **richly described** with a plurality of accurate and relevant attributes
  - (Meta)data are released with a clear and accessible data usage **license**
  - (Meta)data are associated with detailed **provenance**
  - (Meta)data meet domain-relevant **community standards**

euraxess
RESEARCHERS IN MOTION

Difference between FAIR data and Open data

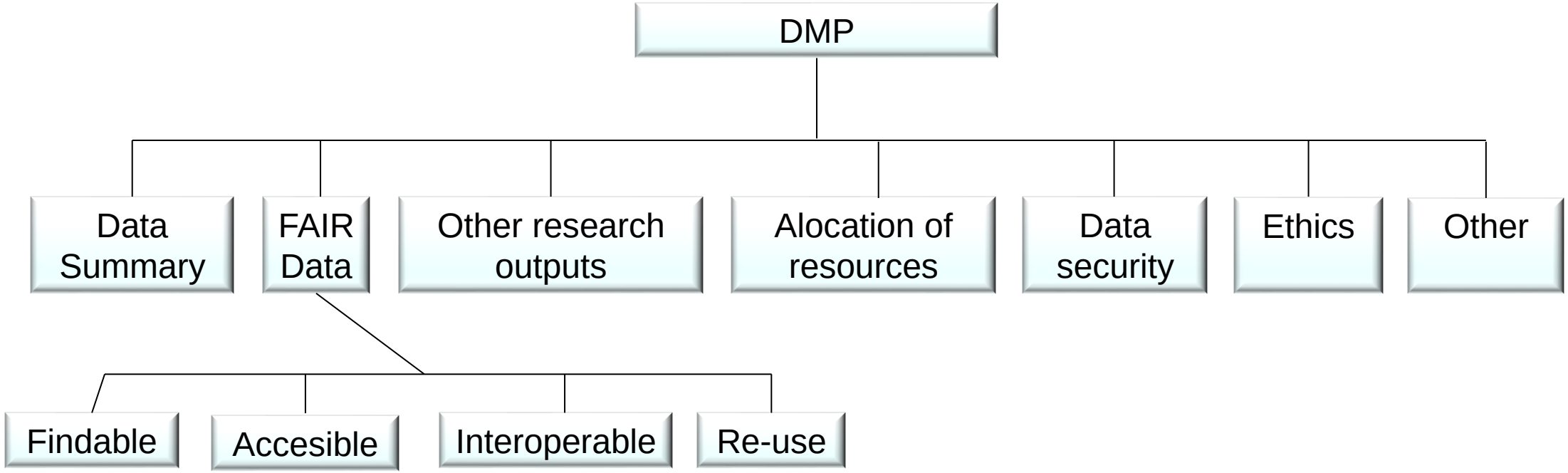
- **Open data** should be **available to everyone to access, use, and share, without licences, copyright, or patents**. At most, it should be subject to attribution/share-alike licenses
- **FAIR data**, uses the term “Accessible” to mean **accessible by appropriate people, at an appropriate time, in an appropriate way**.
- Data can be FAIR when it is private, when it is accessible by a defined group of people, or when it is accessible by everyone (open data).

Data management plan

- A **Data Management Plan (DMP)** is a brief plan to define:
  - how the data will be created
  - how it will be documented
  - who will be able to access it
  - where it will be stored
  - who will back it up
  - whether (and how) it will be shared and preserved
- DMPs are often submitted as part of grant applications, but are useful whenever researchers are creating data
- Further questions: CESSDA Data Management Expert Guide, part 6
- DMP templates across EU: CESSDA European diversity
- Horizon Europe DMP template

euraxess RESEARCHERS IN MOTION

DMP template (Horizon Europe)



euraxess
RESEARCHERS IN MOTION

How to make research data accessible

1. As supplemental material **with a research article**, **hosted by the publisher** of the article
2. Hosting data on a **publicly-available website**, with files available for download
3. Depositing data in a **repository** developed to support data publication (e.g. Zenodo)
4. Publishing **a data paper about the dataset** (as a preprint, in a journal, or in a data journal dedicated to supporting data papers). Hosted by the journal or hosted separately in a data repository

euraxess
RESEARCHERS IN MOTION

Where to publish data?

Order of preference recommended by OpenAIRE:

1. Use an **external data archive** or repository established for your research domain to preserve the data. Some recommendations are given by Nature
2. If available, use an **institutional repository**, or your research group's established data management facilities
3. Use a **cost-free data repository** such as Dataverse, Dryad, figshare or Zenodo.
4. Search for **other data repositories** in re3data. Filter options that will help you find FAIR-compatible repositories: access categories, data usage licenses, trustworthy data repositories and data a persistent identifier (PID). Consider whether the repository supports versioning

euraxess RESEARCHERS IN MOTION

Find a discipline-specific repository

- Open Research Europe lists approved repositories in their Data guidelines chapter 2.2
- List of data repositories categorized by disciplines, from the University of Chigago Library
- List of recommended data repositories by the data journal Scientific data
- Search the re3data registry of research data repositories directly or with the Repository Finder by DataCite
- List of data repositories in Open Access Directory

euraxess RESEARCHERS IN MOTION

Selection of one of the general purpose repositories

- Dataverse
- Dryad
- Fighsare
- Mendeley Data
- OSF
- Vivli
- Zenodo
- Characteristic of general repositories

euraxess RESEARCHERS IN MOTION

Practical steps to publish the research data

- Prepare data for publishing
  - Check if it is allowed to publish data
    - Personal data
    - Confidential data
    - Data to be used commercially
  - Choose which data should be published
    - At minimum, the data that is needed to validate your results
    - Everything that is needed to replicate a study
    - Everything that is potentially useful for scientific community
  - FAIRSharing.org – details of data standards specific to the topic of your research

euraxess RESEARCHERS IN MOTION

Spreadsheet data

DO	DO NOT
Give each column a descriptive heading	Embed charts, comments or tables within a spreadsheet.
Use a single header row	Use color coding (machine-based data mining cannot interpret this)
Ensure you have used the first cell, i.e. A1	Include special (i.e. non alphanumeric) characters within the spreadsheet, including commas
Include a title and a legend to describe each spreadsheet.	Use merged cells
Save each data file with a name that appropriately reflects the content of that file	Deposit multiple worksheets within a spreadsheet (such as those used in Microsoft Excel), as these are not supported by CSV and TAB formats
Deposit each table that is part of the dataset as a separate file	
Deposit each worksheet as a separate file	

euraxess
RESEARCHERS IN MOTION

Practical steps to publish the research data

- Upload data to the repository (choose repository)
  - Persistent identifier: Does the repository provide a DOI or some other identifier?
  - Long-term availability of your data: Is the repository managed by a company, society, institution or government?
  - Impact and visibility: From which repository would the potential reusers best find your dataset?
  - Support for certain metadata, file format or data presentation features. Some discipline-specific repositories have useful features that add value and reuse potential to your data.
  - Choose **certified** repository
  - List of repositories recommended by Open Research Europe

euraxess RESEARCHERS IN MOTION

Practical steps to publish the research data

- Add a **Data Availability Statement** to Your Article
  - All articles must include a Data Availability statement
  - This statement should be added to the end of the article prior to submission.
  - The Data Availability statement should not refer readers or reviewers to contact an author to obtain the data
  - “No data are associated with this article.”
  - “All data underlying the results are available as part of the article and no additional source data are required.”
- **Link your dataset to your article**

euraxess RESEARCHERS IN MOTION

Data citation – Why we need to cite data?

- Benefits for **data producers**:
  - provides proper attribution and credit
  - creates a bibliographic "trail", connecting publications and supporting data
  - demonstrates the impact of their work and establishes research data as an important contribution to the scholarly record
- Benefits for **data users**:
  - citation makes it easier to find datasets
  - supports persistence of datasets
  - encourages the reuse of data for new research questions
- Benefits for **everyone**:
  - increases transparency and reproducibility

euraxess
RESEARCHERS IN MOTION

Data citation services

- **Data citation services** help research communities **discover, identify, and cite research data** (and often other research objects) **with confidence**
- Typically **involves the creation and allocation of Digital Object Identifiers (DOIs) and accompanying metadata** through services such as DataCite (https://www.datacite.org) and can be integrated with research workflows and standards
- An emerging field, that involves:
  - conveying to journal publishers the importance of appropriate data citation in articles,
  - enabling research articles themselves to be linked to any underlying data

Data citation – How it should look

- DataCite format:
- **Creator (PublicationYear). Title. Version. Publisher. ResourceType. Identifier**
- Barclay, Janet Rice (2013) Stream Discharge from Harford, NY. Cornell University Library eCommons Repository. http://hdl.handle.net/1813/34425

Exercise: How to find data

- Google dataset search
- DataCite search
- World Bank Open Data
- WHO (World Health Organisation) Open Data repository
- European Data portal

euraxess
RESEARCHERS IN MOTION

Thank you!

QUESTIONS

misicdr@gmail.com

euraxess
RESEARCHERS IN MOTION