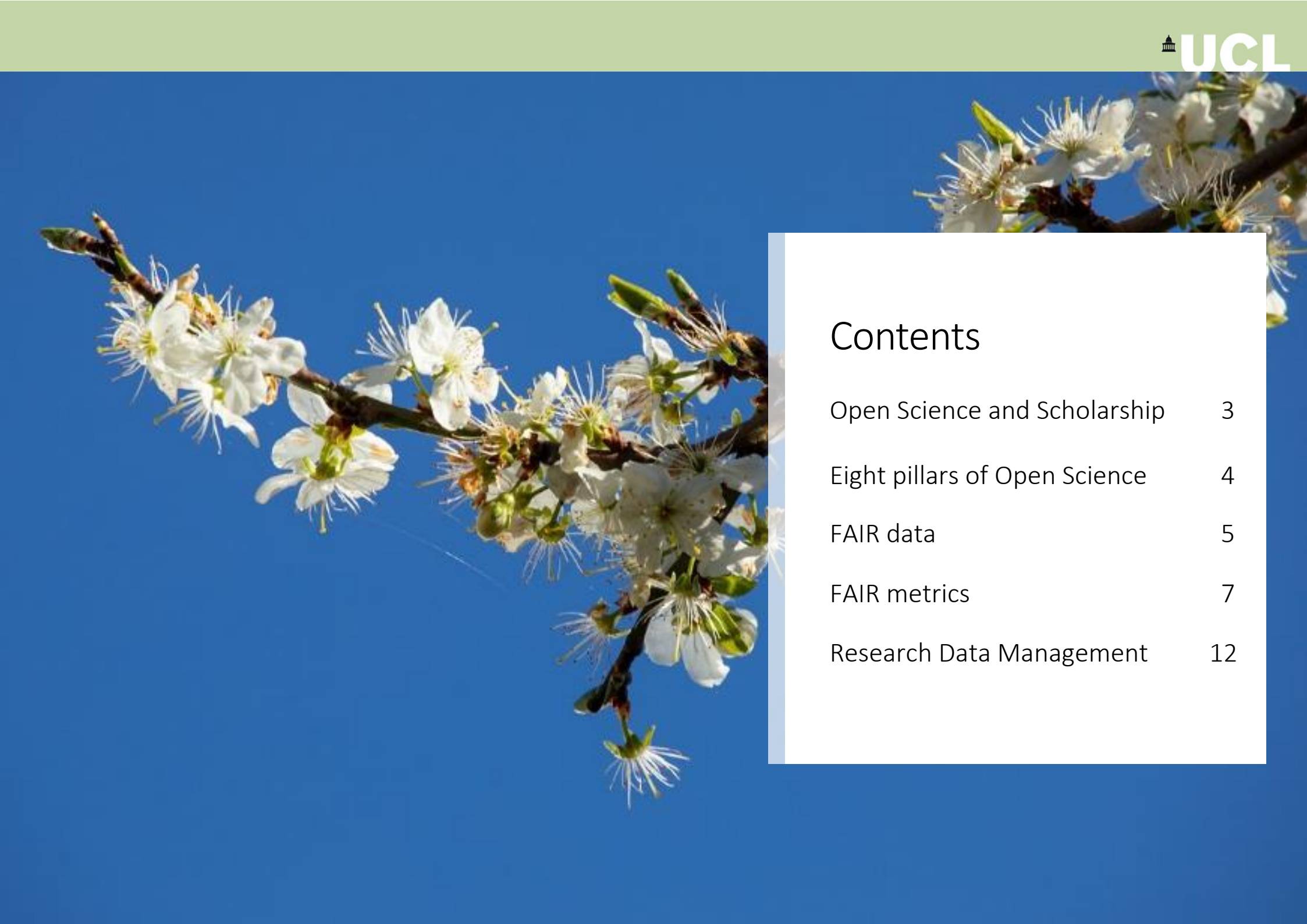
A brief introduction to creating and sharing FAIR data at UCL

Christiana McMahon

James Houghton

Kirsty Wallis



Contents

"Open Science/Open Scholarship provides a set of principles and suggestions for practice which will transform the research landscape across the world to the benefit of Society as a whole[1]*"*

[1] www.ucl.ac.uk/library/research-support/ucl-office-open-science-and-scholarship

Open Science & Scholarship is a new cross-domain approach to creating, sharing, evaluating, rewarding and curating outputs of research[1].

Members of the UCL community are being actively encouraged to embrace open practices to harness the benefits linked with this new and innovative approach, such as:

- **Facilitating greater access to research findings**
- **Promoting research integrity**
- **Maximising research potential of existing resources through reuse and repurposing**
- **Promoting uptake of citizen science**
- **Embedding trustworthiness in your research outputs through greater transparency of the research process**
- **Bringing a greater sense of accountability to research**

The UCL Statement on the importance of Open Science outlines UCL's commitment to tackling the challenges associated with having closed or restricted access to research findings and the overall research process.

Dr Paul Ayris, Pro-Vice-Provost (UCL Library Services & UCL Office for Open Science & Scholarship)**, says: "***For this to work, all players in the research landscape need to embrace Open Science principles and practice to create a new research landscape built on Openness, transparency, sharing and collaboration.***"**

Eight pillars of Open Science[2]

The future of scholarly publishing: the movement towards even greater access to publications by making them as openly accessible as possible. This pillar includes open access publishing[3] initiatives, Plan S and UCL Press[4]

The European Open Science Cloud: European-based, domain-agnostic infrastructure designed to support and quicken the process towards Open Science across Europe ultimately supporting the reuse and repurposing of research data. The vision is to link the EOSC to other international infrastructures to provide a global catalogue of research data.

Education and skills: sufficiently addressing any gaps in knowledge and skills around making publications openly accessible and managing research data appropriately. All researchers at all levels should have access to training programmes to suit their training needs.

Rewards and initiatives: a perceived lack of rewards and recognition for work undertaken to manage research data and make publications openly accessible discourages researchers from engaging with Open Science and Scholarship. Movement away from a dependence on journal impact factors and other traditional measures of success as a proxy for research quality will help to strengthen the argument for greater awareness of the wider Open Science agenda and its role in the future of academic research.

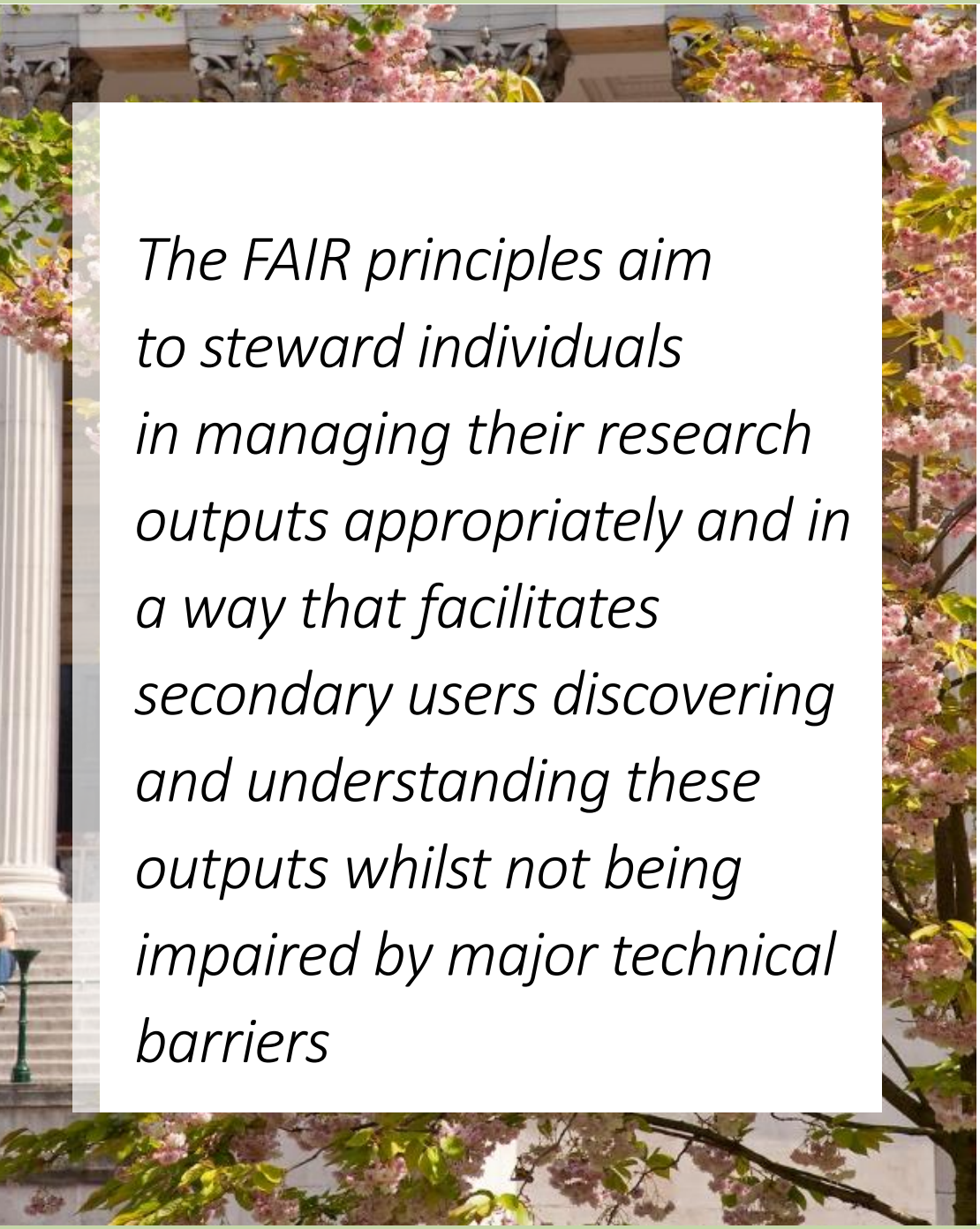
Next-generation metrics: addressing the technical and cultural challenges associated with the way in which traditional bibliometrics are utilised and how metrics may also be applied to other research outputs such as research data which currently fall out of scope of current metrics.

Research integrity – the practice of researchers acting honestly, reliably, respectfully and are accountable for their actions[5]. This is in support of creating a working environment not only conducive to excellent research but also reflective of the wider Open Science agenda.

Citizen science – this is the movement towards members of the public having a greater role within research. Harnessing the advantages of the internet, openly available software packages and local knowledge, citizen science brings about a change in the way research is conducted – no longer confined to universities, it encourages collaboration from groups across society.

FAIR data: facilitating the reuse of data and other research outputs such as software, models, photographs etc. by making these findable, accessible, interoperable and reusable - FAIR.

[2] www.leru.org/files/LERU-AP24-Open-Science-full-paper.pdf
[3] www.ucl.ac.uk/library/research-support/open-access
[4] www.uclpress.co.uk
[5] www.ucl.ac.uk/research/integrity



The FAIR principles aim to steward individuals in managing their research outputs appropriately and in a way that facilitates secondary users discovering and understanding these outputs whilst not being impaired by major technical barriers

Findable

Making research outputs discoverable by the wider scientific community and the public

Accessible

Use of unique identifiers, metadata and use of clear language and access protocols

Interoperable

Applying standardised vocabularies to encode and exchange data/metadata

Reusable

Enabling the repurposing of research outputs to maximise research potential

FAIR data and Open data are not the same thing

The terms FAIR data and Open data should not be used interchangeably - research outputs can be FAIR whilst being subject stringent data access protocols.

Essentially, whilst there could be ethical, legal or commercial constraints on the release of data freely online, this should not stop individuals from making outputs findable, accessible, interoperable and reusable - FAIR.

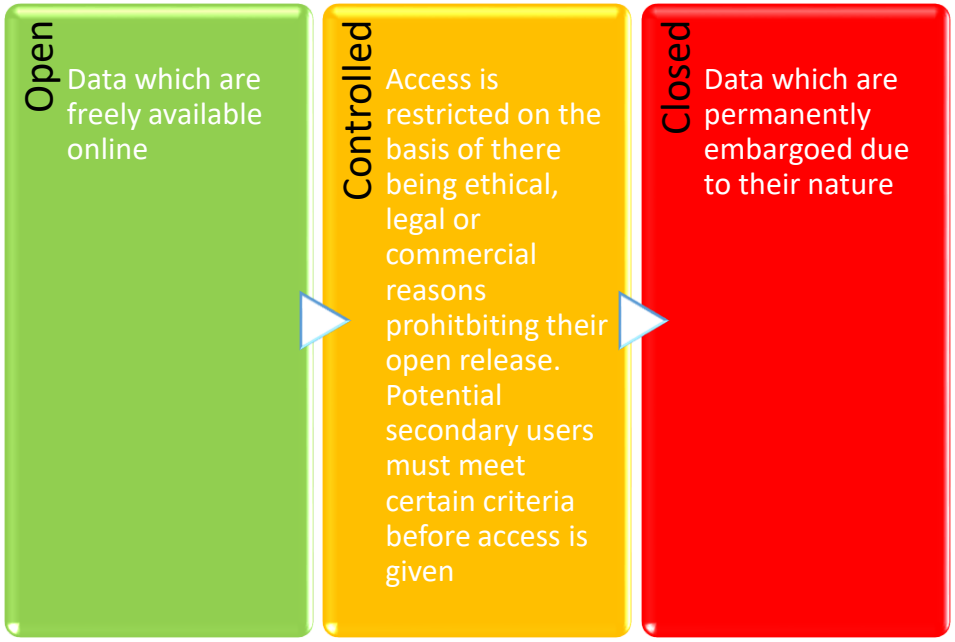


Figure 1: Open vs Controlled vs Closed data

A FAIR metrics[6] guide to creating and sharing FAIR data

Findable

(F1) (META)DATA ARE ASSIGNED A GLOBALLY UNIQUE AND PERSISTENT IDENTIFIER

Digital Object Identifiers, or DOIs, are routinely assigned to research outputs - such as research data - to identify these in a globally unique and persistent way. DOIs resolve to a landing page containing descriptions or metadata.

UCL staff and research students can assign their research outputs a DOI and create metadata using the UCL Research Data Repository[7] (RDR). Individually published records in the RDR can be linked together either as a collection of records, or as project and assigned a DOI.

(F2) DATA ARE DESCRIBED WITH RICH METADATA

Data documentation, or metadata, describe the research data, study materials and the research process across each stage of the research data lifecycle: 1) planning and preparation; 2) actively researching; 3) archiving, preserving and curating data; and 4) data discovery, access and sharing. Metadata can provide potential secondary users with contextual information to facilitate the reuse and repurposing of existing research outputs helping to maximise their research potential.

UCL staff and students should begin creating metadata at the start of the research project and regularly document the study as it progresses. Researchers should aim to apply a metadata standard[8] and meet community expectations, where possible.

(F3) METADATA CLEARLY AND EXPLICITLY INCLUDE THE IDENTIFIER OF THE DATA IT DESCRIBES

Published data can be assigned a unique identifier e.g. a DOI, as described in **F1**. This unique identifier should be included in the accompanying metadata to describe where the data are. This DOI points to a landing page that describes the published item(s), any restrictions on data access if applicable and if there are protocols in place to govern access to the data.

> Katsianis, Markos; Bevan, Andrew; Styliaras, Giorgos; Maniatis, Yannis (2020): Dataset for: An Aegean history and archaeology written through radiocarbon dates. University College London. Dataset. https://doi.org/10.5522/04/12489137.v1

Figure 2 Example of a data citation with DOI link

[6] Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific Data, 2016. **3**: p. 160018

[7] www.ucl.ac.uk/library/research-support/research-data-management/ucl-research-data-repository

[8] www.dcc.ac.uk/drupal/resources/metadata-standards

(F4) YOUR METADATA AND DATA ARE REGISTERED OR INDEXED IN A SEARCHABLE RESOURCE

These resources tend to be online, publicly accessible repositories or archives that allow individuals to create metadata records - with a unique persistent identifier such as a DOI – and preserve data for the medium to longer-term. These metadata records tend to describe the research project, the item(s) deposited and how the item(s) may be accessed and reused, if possible. Having a publicly accessible record of the research can help to enhance the discoverability of the research study and potential for reuse.

Accessible

(A1) (META)DATA ARE RETRIEVABLE BY THEIR IDENTIFIER USING A STANDARDIZED COMMUNICATIONS PROTOCOL

This metric refers to the technical and organisational mechanisms used by organisations to create and offer unique, persistent identifiers for research outputs. For example, if a DOI has been assigned, it will follow a standardised, internationally recognised protocol as defined by the DataCite[9] metadata schema. UCL staff and research students should aim to assign unique identifiers that utilise standardised protocols wherever possible.

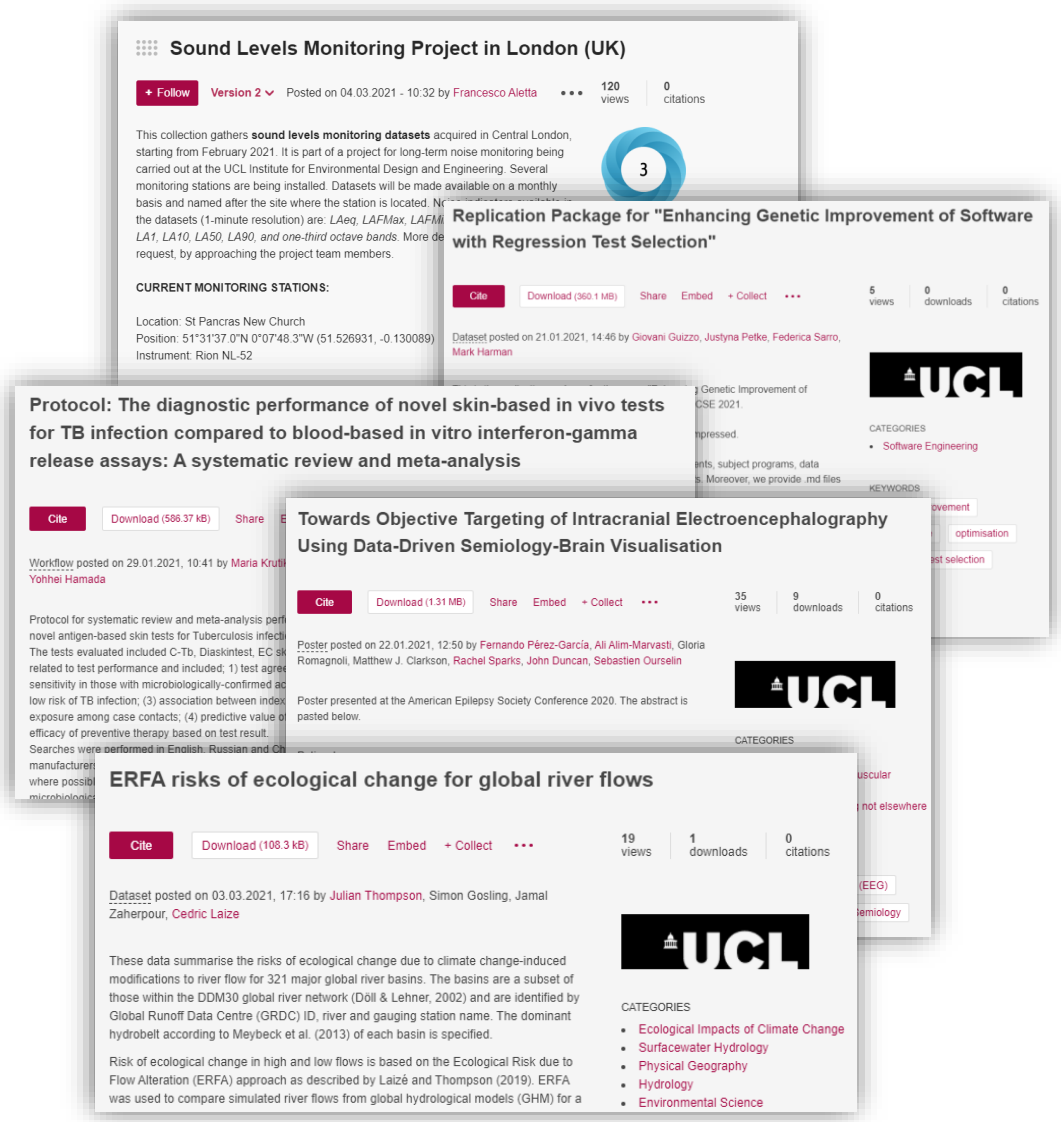


Figure 3 Examples of published records indexed in the UCL Research Data Repository

[9] https://datacite.org

(A1.1) THE PROTOCOL IS OPEN, FREE, AND UNIVERSALLY IMPLEMENTABLE

Whichever unique identifier a researcher selects to use, the protocol discussed in **A1** should not be subject to any major technical, organisational or financial barriers. Individuals globally should be able to implement the protocol relatively easily facilitating efficient access to the research outputs.

(A1.2) THE PROTOCOL ALLOWS FOR AN AUTHENTICATION AND AUTHORIZATION PROCEDURE, WHERE NECESSARY

Authentication and authorisation may be needed when trying to access data that are subject to a set of controls or restrictions. Authentication is the process of verifying the person trying to access the research output. Authorisation refers to the individual permissions settings the person is subject to once access has been granted. Individuals should be mindful of any ethical, legal, and/or information governance frameworks which could inform decision making around use of authentication and/or authorisation procedures.

(A2) METADATA ARE ACCESSIBLE, EVEN WHEN THE DATA ARE NO LONGER AVAILABLE

If research outputs are no longer available, as they were destroyed following approval to do so, or are subject to ethical, legal, commercial and/or information restrictions rendering them 'closed', individuals should continue to provide and maintain metadata for these data. This is in-line with best practice and in certain cases, mandated by stakeholders such as funders. Those responsible for maintain the metadata should include a description covering why the data are no longer available.

Interoperable

(I1) (META)DATA USE A FORMAL, ACCESSIBLE, SHARED, AND BROADLY APPLICABLE LANGUAGE FOR KNOWLEDGE REPRESENTATION

Individuals should create study documentation (metadata) using clear and accessible language. Where applicable, individuals should aim to embed metadata in software and other coding through use of labelling and commentary. Additionally, if there are community standards for encoding information, such as ICD codes for medical/clinical information, the metadata should utilise these wherever possible to create standardised and widely understood metadata.

Where possible, individuals should create a machine-readable version of the metadata such as .XML. In certain cases, some online repositories and digital archives can automatically create machine-readable versions of the metadata.

(I2) (META)DATA USE VOCABULARIES THAT FOLLOW FAIR PRINCIPLES

Individuals should aim to apply community-accepted/approved controlled vocabularies which are implementable without any major technical barriers and in-line with the FAIR principles more widely. Vocabularies and other such community standards which are in-line with the FAIR principles can be found online at FAIR Sharing[10].

[10] https://fairsharing.org

I3. (META)DATA INCLUDE QUALIFIED REFERENCES TO OTHER (META)DATA

Building an online profile of your research by creating a network of linked outputs requires individuals make use of unique identifiers such as DOIs and ORCIDs.

Individuals can also assign a DOI to a metadata record in an online catalogue for their deposited data and study materials, and then include this DOI in the data access statement in the associated publication, as discussed. This complements the advice given in **F3**.

Reusable

(R1) META(DATA) ARE RICHLY DESCRIBED WITH A PLURALITY OF ACCURATE AND RELEVANT ATTRIBUTES

Metadata are descriptions of data and other research outputs over the course of the research study. Examples of metadata include data dictionaries, codebooks, readme files, blank questionnaires and consent forms etc. The different kinds of metadata are: administrative – describes what is needed to open, read and the use data; descriptive - describes what is needed to understand and reuse the data; and semantic – describes how the different aspects of the data relate to each other.

Individuals should create metadata in real-time using tools to help automate the process as much as possible. Creating metadata retrospectively is possible, but there is the potential for recall bias.

[11] www.ucl.ac.uk/library/research-support/research-data-management/best-practices/how-guides/software-sustainability

[12] www.ucl.ac.uk/library/ucl-copyright-advice

(R1.1) (META)DATA ARE RELEASED WITH A CLEAR AND ACCESSIBLE DATA USAGE LICENSE

A usage license instructs other (potential) users of your research output(s) others on how they may been used[11]. Creative Commons licenses are routinely utilised in academia; CC-BY (the attribution-only license) is commonly assigned to publications and research data.

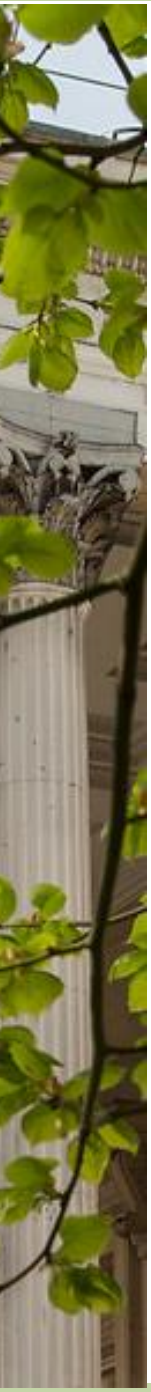
UCL Library Services provides a copyright advisory service[12] to support individuals in matters relating to copyright. Where staff and students wish to explore the (potential) commercial value in their research outputs - and look to protect their IP[13] and other rights – they should speak with UCL Business[14].

CC0 - No rights reserved	Explanation
PUBLIC DOMAIN	CC0 goes one step further than the Creative Commons licences in making your work available for others to reuse. CC0 enables you to waive any claim you might have to copyright protection in the work. It follows that others will be free to reuse your work in any way they choose without seeking permission, without acknowledging you as the author and with no other formalities. If the CC licences are equivalent to "some rights reserved," CC0 is equivalent to "no rights reserved."

Figure 4 Creative Commons Licenses I
www.ucl.ac.uk/library/ucl-copyright-advice/creative-commons-licences

[13] www.ucl.ac.uk/enterprise/about/governance-and-policies/ucl-intellectual-property-ip-policy

[14] www.uclb.com



CC Licence	Explanation
CC BY	You can reuse the work in any context including for commercial purposes, such as reproducing it in a commercial publication or placing it behind a pay wall. You can produce a derivative version, such as a mash-up of multiple images, an animated version of a novel or a translation into another language.
CC BY-NC	Like CC BY with the additional restriction that the work cannot be reused for commercial purposes.
CC BY-ND	Like CC BY with the additional restriction that you cannot make derivative works available. It can only be reused in its original form, so for example no mash-ups, translations, simplified versions or adaptations for a different medium.
CC BY-SA	The addition of "SA" means that if you make a derivative version available, it must be published under exactly the same licence as the original, and without placing any additional restrictions.
CC BY-NC-SA	Permits derivative versions but they can only be made available under the CC BY-NC-SA licence. NC means that the licence does not allow any commercial reuse.
CC BY-NC-ND	Attribution- Non-commercial - No derivatives. This is the most restrictive CC licence. It does not permit commercial reuse or the making available of derivative versions.

Figure 5 Creative Commons Licenses II
www.ucl.ac.uk/library/ucl-copyright-advice/creative-commons-licences

(R1.2) (META)DATA ARE ASSOCIATED WITH DETAILED PROVENANCE

Provenance refers to the history of the research output and is often included in the metadata. For example, if you entered into a data sharing agreement with a collaborator, staff and students should record that data were shared, by a named organisation and under certain terms and conditions. Any subsequent processing of that data including moving the outputs into an archive or repository should be recorded in the provenance.

(R1.3) (META)DATA MEET DOMAIN-RELEVANT COMMUNITY STANDARDS

Staff and students should apply a domain-specific metadata standard[15] to help create documentation that meets community expectations, where possible. Individuals can apply generic metdata standards too[16].

Also, staff and students should standardise the format and structure of the research outputs too. For example, by standardising research datasets, any secondary users can potentially have greater opportunity to understand the data helping to maximise their research potential through reuse. This metric complements that discussed in **I2**.

[15] https://fairsharing.org

[16] www.dcc.ac.uk/drupal/resources/metadata-standards

The Research Data Management team can help the UCL community to create and share FAIR data

The RDM team can advise on:

- **Writing data management and sharing plans**
- **Meeting UCL and funder's RDM expectations**
- **Adopting open research practices**
- **Archiving items using the UCL Research Data Repository**
- **Storing and securing research data and study materials**
- **Archiving, preserving and curating research outputs**
- **Facilitating the discovery, access and sharing of data**

Contact us

Email: lib-researchsupport@ucl.ac.uk
Website: https://bit.ly/3rgFe0l
Twitter: @UCLopenscience