METHODS USED TO ADAPT FAIR PRINCIPLES TO SCIENTIFIC DATA

Cristian CUDALBU, Teodora BABOS, Florin PITEA

National Institute of Marine Geology and Geo-Ecology (GeoEcoMar), 23-25 Dimitrie Onciul St., 024053 Bucharest, Romania
e-mail: c.cudalbu@geoecomar.ro, teodora.babos@geoecomar.ro, florin.pitea@geoecomar.ro

DOI: 10.xxxx

Abstract. This paper presents an analysis of the FAIR principles. It is addressed to data managers and scientists (as data producers and users) and aims to point out the crucial importance of providing proper metadata and take all the necessary actions in order for the databases to follow the FAIR principles.

Key words: FAIR, digital object, research data, data management

1. INTRODUCTION

In the recent years due to technological progress there is a very large amount of scientific data generated from research activities, but in many cases, it is not possible - and certainly not easy - to find this data and query it to compare it with other recent discoveries or work in progress. One of the most recent and prominent approaches to solving this problem is FAIR, which encompasses a set of principles about how data should be organized to be as usable and valuable as possible: FAIR data is data that is „Findable"; „Accessible"; „Interoperable" and „Re-usable":

- "**Findable**" means that scientific data that is published as part of the evidence underlying scientific discoveries or produced as a result of publicly funded research should be available for others to find and use. Data should have a persistent and unambiguous identifier, as well as metadata rich enough to enable its discovery.
- "**Accessible**" means that it must be possible to access the data in its entirety online, via the web, with authorization services if there are security issues. Essentially, the FAIR principles argue that scientists should be able to access research data programmatically, computers also. Accessible does not just mean that the data can be obtained and downloaded: ideally it should be able to be queried using dedicated programming languages and scripts.
- "**Interoperable**" means that data from different sources can be combined - this depends largely on the existence of metadata standards and agreed terminologies or vocabularies.
- "**Reusable**" – this term defines the inclusion of a license that allows people to reuse the data and clearly states the conditions for any reuse. It also means having information about the provenance of the data (how it was collected, what adjustments or calibrations were used, what post-processing the data underwent, etc.) so that researchers can understand potential strengths and limitations of the data in order to use it with confidence.

FAIR Data is also "Fully AI Ready": for machine learning neural networks that identify patterns and predict outcomes in different data sets, it is essential to have definitions for different variables in the data set, and the definitions must be easily accessible.

More importantly, the associated metadata provides value as a permanent scientific record, even when the original data have lost their immediate value so that they are archived or eventually deleted as part of the FAIR data management lifecycle.

It should be noted that FAIR data is not the same as „open" and „open source" data. All these initiatives aim to remove practical obstacles for the use of information. In the case of FAIR, it is about **compatibility issues, and not about legal issues**.

„Open" data aims to make data available to the public for free. This removes financial barriers to data reuse and greatly facilitates innovation.

Publishing data as „open" data does not make the data perfect: if the data is not in the correct format or descriptions are missing, the data may be „open" but still difficult to use, and it could even produce false information as the result of the wrong interpretation of the primary data.

Currently, most of the FAIR data are data resulting from scientific research activity and are also „open".

Many of the FAIR principles depend on the availability of metadata. Metadata means the data „about the data," and the word is used to refer to the explanation and documentation you need to understand the data. The main problem with classic data sharing is that context/metadata is often lost in transit or it is not produced according with the accepted standards. FAIR data means data with the right metadata to make the data clear and useful.

FAIR has a „complicated" relationship with privacy. GDPR prohibits data sharing between organizations without a clear purpose. FAIR data principles could also apply to personal data, however, it is unlikely that this data will become freely available.

FAIR is already quite widely used in the natural sciences. Other sectors are expected to follow. The FAIR principles are domain independent, so they can be applied to any domain: from log files to CRM and financial data to maps and images.

2. ANALYSIS

In documenting data with metadata, it is important to follow standard vocabularies, standard schemas, etc. agreed by professionals acting in a specific domain. Organizing data using a standards-based approach helps ensure interoperability between systems, which also improves data discovery and access.

The following analysis was made for scientific data on the four principles of FAIR following the most relevant standards and schemes used, aiming to adapt the solutions for scientific data in order to obtain the most appropriate „Data Fairness".

2.1. Accessibility and interoperability – the need for a DOI for data

Globally unique and persistent identifiers remove ambiguity in the meaning of published data by assigning a globally unique identifier to each metadata element and each concept/measurement in the dataset. In this context, identifiers consist of an Internet link (*e.g.*, a URL that resolves to a web page that defines the concept). Many data repositories will automatically generate globally unique and persistent identifiers for stored datasets. Identifiers can help other people understand exactly what we mean and allow computers to interpret the data in a meaningful way (*i.e.* computers searching for the data or trying to integrate it automatically). Identifiers are essential for human-machine interoperability, which is key to the vision of Open Science. Additionally, identifiers will help others properly cite our work when they reuse the data.

To ensure the validity of a unique identifier, two conditions must be met:

- Must be globally unique (*i.e.* someone else could not reuse/reassign the same identifier without referring to our data). We may obtain globally unique identifiers from a registry service that uses algorithms that guarantee the uniqueness of generated identifiers (DataCite obtaining a DOI case study is described below).
- It must be persistent. It takes time and money to keep web links active, so links tend to become invalid over time. Registry services guarantee the resolvability of that link in the future, at least to some extent (FAIR Principles, https://www.go-fair.org/fair-principles/).

2.1.1. Obtaining a DOI with DataCite

One of the most well-known organizations in the scientific community for registering persistent identifiers is DataCite.

DataCite is a leading global non-profit organization that provides persistent identifiers (DOIs) for research data and other research outputs. Organizations in the research community join DataCite as members to be able to assign DOIs to all of their research results. In this way, their results become discoverable and the associated metadata is made available to the community. DataCite then develops additional services to enhance the DOI management experience by making it easier for existing members to connect and share their DOI with the broader research ecosystem and to evaluate the use of their DOIs within that ecosystem. DataCite is an active participant in the research community and promotes data sharing and citation through community building efforts and outreach (https://datacite.org/).

DataCite offers the following facilities to its members:

- Metadata recorded with DataCite resides in a central location that can be harvested by anyone.
- Metadata for member search results appear in other search engines.
- DOIs and DataCite metadata increase research fairness.
- Connection with the community of DataCite members, offering the opportunity to exchange experience between them.
- The metadata scheme is extensive and has been adopted by other PID service providers globally.
- DataCite services simplify institutional reporting.
- DataCite services support data citation and usage analysis.

"Fabrica" is DataCite's web interface where all DOIs and metadata can be created, found, linked and tracked. The factory also includes all the necessary functionality to manage deposit accounts, prefixes and contacts.

In order to create a DOI, after registering for an account with DatCite from the "Fabrica interface" the following steps must be taken:

Preparing the data repository

Before submitting the DOI request, we need to ensure that all files have been uploaded and that the repository contains a comprehensive description of the contents, so that other researchers can understand and access our data or code.

The repository name

We must choose a name for our repository that is brief, but specific. If a user downloads the repository or its archive from the server, its contents will be put in a folder named after the repository name. Therefore, the repository name should be unique and descriptive of our data. We should avoid generic names (like „my_data", „plos_paper", or the like). If we create a repository specifically for a supplement to a paper publication, a name of the form <first_author>_et_al_<year>_<journal> may be useful to consider.

The README file

The usual place to present information about the repository for the user is the README.md file at the root of our repository. Depending on the structure and contents of the repository, it may be useful to include further README files in subfolders.

The description of our repository should include:

- A title and summary of what the dataset is about.
- Information about the repository content and structure.
- Information on how to access and use the data and/or code, including what data formats are used, how metadata is provided, what code is available and how to use it, etc.

It is useful to also include in the README file general information, like authors, contact information, acknowledgments, links and references, etc. Since the DataCite information is for automated creation and registration of the DOI record, while the README file is intended for human readers, some redundancy between these files is not a problem.

An important part of accessibility is the definition of usage licenses: as mentioned in the introduction FAIR data does not equal free access data. These licenses should be both human and machine readable, clear in their limitations and user rights and accessible online in all cases.

Also, it should be noted that even if data access is not entirely free access to metadata should be totally free. Metadata should also be available many years after the data lost its relevance in order to increase the accessibility to data.

Creating a DOI

The "Create DOI" button appears on the left-hand side of all of the tabs on the Repository dashboard. Click the Create DOI button to go directly to the Form or hover over the Create DOI button and click DOI Form to register a new DOI using the Form (Figure 1).

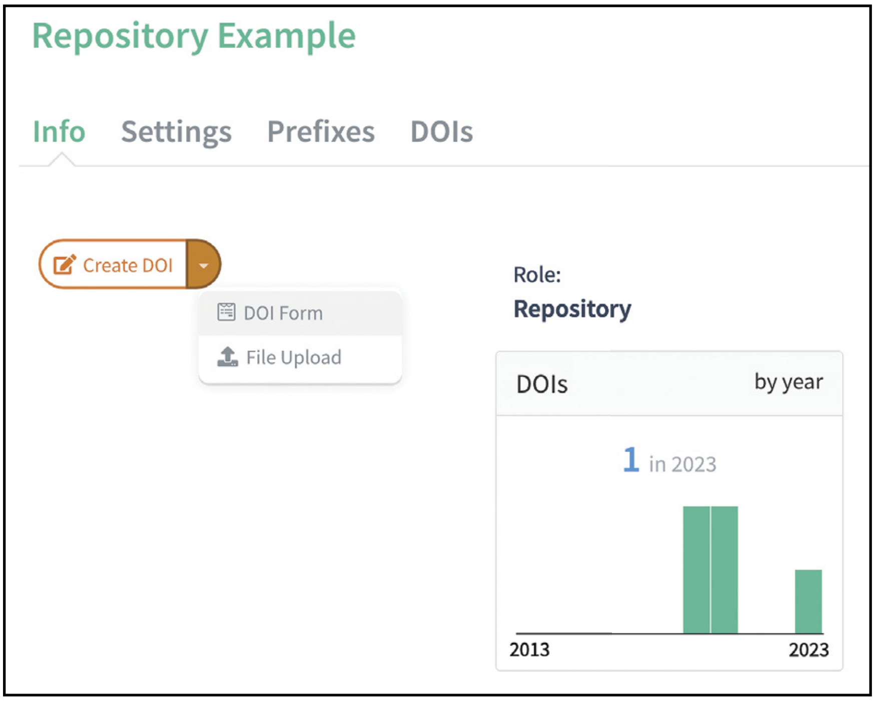


Fig. 1. Form for creating a new DOI.

The DOI name appears at the top of the page. The suffix is automatically generated (this is the recommended option, but it can be edited). We select the state of our data and then enter the metadata. Draft state does not require metadata, only the identifier and the state (selected as Draft). Registered and Findable DOIs need to have all the required metadata fields. The fields in the Form correspond to the DataCite Metadata Schema (Figure 2).

Once we have filled in the appropriate fields, we click the Create DOI button at the bottom of the page. Our DOI will be saved in Fabrica, and the record page for the DOI will be displayed.

DOI records can be created in the Datacite Fabrica interface or using the DataCite API can be created in a self-developed data portal solution or using an existing open source one – CKAN, Geonetwork – to name just two very used data portal solutions – have already implemented endpoints and plugins for DataCite DOI generation

While DataCite is one of the most used services for sceitific data, there are other solution for particular fields of reserch like "SeaDataNet DOI minting service" for publishing marine research data, da|ra - DOI Registration Service for social science and economic data, etc

2.2. Interoperability and reuse

One of the goals of implementing FAIR principles is achieving "semantic interoperability": ability of machines to exchange (use. reuse) data between them without human intervention. For the machines in order to understand data it needs to be accompanied by metadata. There are other types of interoperability that are defined in FAIR principles like legal and organizational (ensuring that organizations operating under different legal frameworks, policies and strategies are able to work together), technical (which covers the applications and infrastructures linking systems and services. Aspects of technical interoperability include interface specifications, interconnection services, data integration services, data presentation and exchange, and secure communication protocols.)

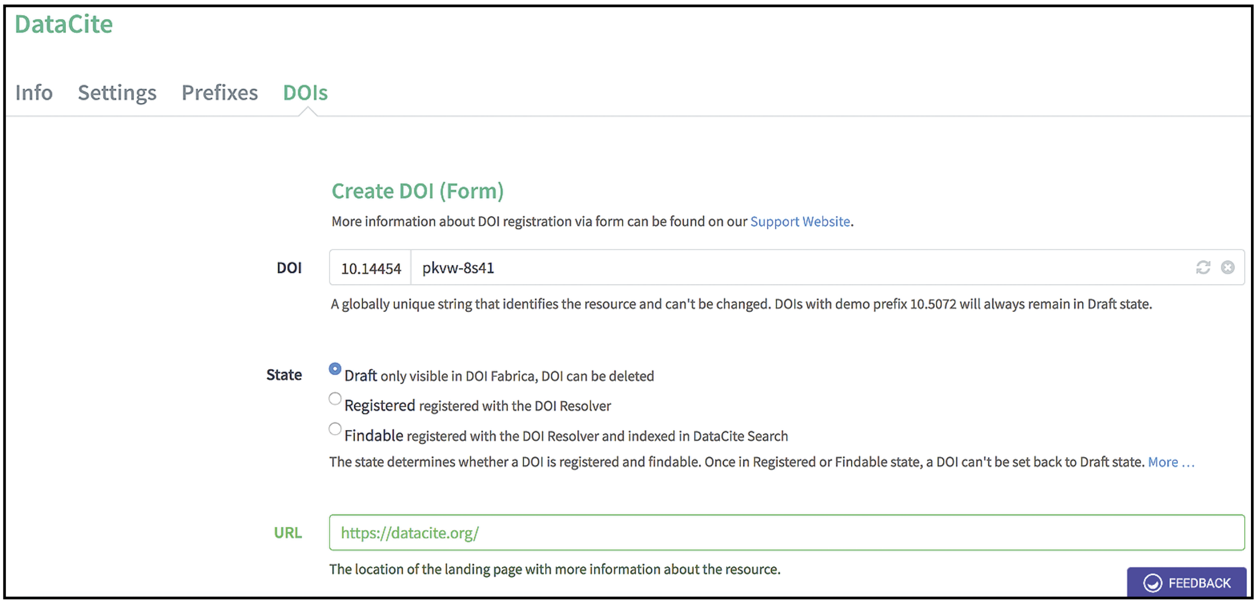
Still, semantic interoperability is the most important from the scientific content point of view.

Interoperability for our scientific datasets can be achieved at the level of the dataset metadata and at the level of the data: good dataset metadata can make a dataset discoverable; good metadata about the data structure can make the data parsable (in addition, using semantics such as standardized variable names linked to agreed vocabularies or controlled values in the actual data records will make data understandable and fully interoperable (GODAN Action Open Data Management in Agriculture and Nutrition Course, 2020).

It is important to be aware of the fact that there are different layers of metadata as displayed in Table 1.

While Catalog, DataSet, Distribution and Data record are straightforward in their meaning, FDP needs further explanation: a FAIR Data Point (sometimes abbreviated to FDP – Fair Data Point) ultimately stores information about datasets (the general definition of metadata). And just as the WWW web server in the early 1990s brought anyone the power to publish text, a FAIR data point aims to give anyone the power to put their own data on the web.

The system is called a FAIR data point because it takes care of many of the issues that need to be taken care of in order for the data to be fair; in particular with the necessary metadata for discovery and reuse and a uniform open way of accessing the data.



Create DOI (Form)

More information about DOI registration via form can be found on our Support Website.

DOI: 10.14454 / pkvw-8s41

A globally unique string that identifies the resource and can't be changed. DOIs with demo prefix 10.5072 will always remain in Draft state.

State:
● Draft only visible in DOI Fabrica, DOI can be deleted
○ Registered registered with the DOI Resolver
○ Findable registered with the DOI Resolver and indexed in DataCite Search

The state determines whether a DOI is registered and findable. Once in Registered or Findable state, a DOI can't be set back to Draft state. More ...

URL: https://datacite.org/

The location of the landing page with more information about the resource.

Fig. 2. DataCite Metadata Schema.

Table 1. Metadata layers (Luiz Olavo Bonino *et al.*, 2016, 'Fair Data Technology Update').

Layer	Description	Example
FDP (Data repository)	Information about the FDP as a data repository	PID, title, description, license, owner, API version, etc.
Catalog	Information about the catalog of datasets offered	PID, title, description, publisher, etc.
Dataset	Information about each of the offered datasets	Publisher, issue date, theme, etc.
Distribution	Information about how the dataset is distributed	AccessURL, downloadURL, format, mediaType, etc.
Data record	Information about the actual data, types, identifiers, etc.	Data items types, identifiers, domain, range, etc.

The FAIR data point also addresses the interoperability of the metadata it stores, but leaves the interoperability aspects for the data itself up to the data provider.

The ultimate goal of FAIR is to optimize the reuse of data. In order to achieve this goal documentation of the datasets should provide a minimum of richness of the metadata. This can be achieved by attaching a simple text file to the dataset archive in a portal for example. This text file should contain the data responsible, what instrument/ programming tool was used to generate the data, methods and procedures used in processing of data if the data is not in raw format, when and where was data generated, comments, etc. as documented in the subchapter 2.1.1 - *Obtaining a DOI with DataCite – The readme file*.

2.3. Findability

According to FAIR principles for data to be findable it needs to:

- Have globally unique and eternally persistent identifier for metadata (which was discussed before in subchapter 2.1);
- described with rich metadata (the level of "richness" or more precise the threshold for the richness of metadata is not specified);
- (meta)data are registered or indexed in a searchable resource;
- metadata specify the data identifier.

In order to achieve these requirements, the best way is to define a metadata schema based on the ones already existing. The most used one in the scientific data field is "Dublin Core".

The Dublin Core™ Metadata Initiative, or „DCMI", is an organization that supports innovation in metadata design and best practices in the metadata ecology. DCMI operates openly and is supported by a paid membership model.

The Dublin Core™ Metadata Initiative (DCMI) supports joint innovation in metadata design and best practices across a wide range of purposes and business models.

DCMI does this by:

- Managing the long-term maintenance and development of metadata term specifications and namespaces;
- organization of an annual international conference;
- access and open availability of meeting assets, including proceedings, project reports and meeting minutes;
- create and deliver training resources in metadata best practices, including tutorials, webinars and workshops;
- coordination of the global DCMI volunteer community.

The operating principles of DCMI are:

- Open Consensus Building: Participation in the DCMI community is open to all groups or individuals interested or experienced in metadata. DCMI de facto standards, specifications and best practice documents reflect consensus reached through consultative debates and reviews. No fees are charged for the use of this information to the extent that the value of these materials is enhanced by their widespread adoption.
- International scope and participation: DCMI emerged in the 1990s from a series of informal workshops that attracted the participation of a worldwide community. DCMI has been committed to global participation from the beginning, as exemplified by the wide range of translations, the location of the Dublin Core™ conferences, and the diversity of regional representation among DCMI members.
- Neutrality of purposes and business models: DCMI is neutral regarding the purposes for which DCMI metadata standards and specifications can de facto be used. DCMI encourages the adoption of these standards and specifications in the public and private sectors and the continuation of de jure standardization that does not jeopardize open access.
- Technology Neutrality: The de facto DCMI standards are fundamentally concerned with semantics - the meaning of statements about information resources. The technological infrastructure underlying the encoding and expression of these semantics is expected to evolve over time. DCMI attempts to maintain the independence of agreed semantics and to facilitate the expression of these semantics by encoding idioms appropriate to the initiative's active stakeholders.
- Interdisciplinary focus: Since its inception in the mid-1990s, DCMI's founding principle has been the discovery and management of metadata resources across the boundaries of web-based information repositories within intranets (About DCMI, 2024).

DCMI metadata terms are expressed in RDF vocabularies for use in the„semantic web" framework.

RDF "triples" with OWL descriptors and DCAT are the main pillars of semantic web. An RDF table has three columns (hence the tern "triples"): subject, verb, and object. There are no relations between RDF triples like in a classic relation database. RDF are much easier to query than relational databases using a dedicated SQL syntax: SPARQL.

There is a common misconception that RDF is open source when in fact parts are patent owned by Oracle who choose not to exercise their patent rights in the present.

Dublin core can also be used for non-semantic databases; non-RDF metadata creators can use terms in contexts such as XML, JSON, UML, or relational databases, ignoring both the global identifier and the formal implications of the RDF-specific aspects of term definitions (non-semantic web).

Each term is identified with a Uniform Resource Identifier (URI), a global identifier that can be used in „Linked Data". Term URIs resolve to the document (DCMI Metadata Terms) when selected in a browser or, when referenced programmatically by RDF applications, to one of the four RDF schemas. The scope of each RDF schema corresponds to a „DCMI namespace" or a set of DCMI metadata terms that are identified using a common base URI as listed in the DCMI Namespace Policy. In Linked Data, URIs for DCMI namespaces are often declared as prefixes to make data, queries, and schemas more concise and readable.

The Dublin Core™ Metadata Initiative provides access to schemas that define DCMI term statements represented in various schema languages. Schemas are machine-processable specifications that define the structure and syntax of metadata specifications in a formal schema language.

Analogous to those discussed above related to web semantics, non-semantic web (XLMS) or semantic web (RDFS) schemes can be used (XMLS - XML Schema)

XML Schemas provide a means of defining the structure of XML documents, including metadata. XML Schema is a specification developed and maintained under the auspices of the World Wide Web Consortium.

Built into the Dublin Core standard are definitions of each metadata element – like a native content standard – that state what kinds of information should be recorded where and how. Associated with many of the data elements are data value standards such as the DCMI Type Vocabulary and ISO 639 language codes, etc. More information can be found on the Dublin Core Metadata Initiative website (https://www.dublincore.org/) (Table 2).

One can use these predefined elements like a metadata schema for their data or add new ones in order to achieve a necessary level of richness for their metadata in order to increase findability of data as much as possible.

Another widely used method to increase findability of data is use of controlled vocabularies (*e.g.* DCAT).

Table 2. Dublin Core Metadata Element Set.

Dublin Core Element	Use
Title	A name given to the resource.
Subject	The topic of the resource.
Description	An account of the resource.
Creator	An entity primarily responsible for making the resource.
Publisher	An entity responsible for making the resource available.
Contributor	An entity responsible for making contributions to the resource.
Date	A point or period of time associated with an event in the lifecycle of the resource.
Type	The nature or genre of the resource.
Format	The file format, physical medium, or dimensions of the resource.
Identifier	An unambiguous reference to the resource within a given context.
Source	A related resource from which the described resource is derived.
Language	A language of the resource.
Relation	A related resource.
Coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.
Rights	Information about rights held in and over the resource.

By using community-consented gold standard vocabularies and a formal knowledge description language, findability and interoperability according to the FAIR principles are significantly improved.

DCAT is a vocabulary for publishing data catalogs on the Web, which was originally developed in the context of government data catalogs such as data.gov and data.gov.uk, but is also applicable and can be used in other contexts.

3. CONCLUSIONS

FAIR principles are just that: a set of guidelines for achieving better collaboration and disseminations of data. There is no "FAIR standard" yet, and it depends on each organization how it concludes that their data is FAIR. Nevertheless, from this article we can synthetize the main methods needed to achieve a high degree of data fairness presented in Table 3.

Tabel 3. Main methods needed to achieve a high degree of data fairness.

FAIR Data Criterion	Methods developed to fulfill the criterion
"findable"	- Generation of DOI (digital object identifier) for datasets - Data description using the most complete catalog of metadata using definitions from existing schemes and analyzed in this phase: DataCite, Dublin Core, DCAT
"accessible"	- Use of DOIs for datasets - Creation of access licenses to data sets and a registration mechanism and user rights as well regulated as possible - Free access to metadata even if the data is inaccessible
"interoperable"	-The use of relevant standards for the descriptive metadata of the data sets that will be stored in the portal - use of controlled vocabularies, dictionaries, metadata schema
"reusable"	Complete documentation of datasets: tools used, data author, procedures, and access licenses.

ACKNOWLEDGMENT

The author wishes to express his special thanks to Dr. **Mihaela Melinte-Dobrinescu** for providing literature and useful comments on an earlier version of the manuscript. Also, I wish to express my gratitude to Dr. **Gabriel Ion** for helping review this article before its final form.

The financial support of this paper was provided by the Ministry of Research, Innovation and Digitization through Program1 – Development of the national research-development system, project PFE 23/30.12.2022 AMBIACVA, Program Nucleu – GEOMARDIGITAL, Objective 4 "New environmental analysis tools and methodologies, coupled with free access to data for the transition to open science", Project 23300402 – "INCD GeoEcoMar database organize for the operationalization of the Fair Data concept.

REFERENCES

FAIR Principles, https://www.go-fair.org/fair-principles/

DataCite, https://datacite.org/

GODAN Action Open Data Management in Agriculture and Nutrition Course, https://aims.gitbook.io/open-data-mooc/, 2020

Luiz Olavo Bonino *et al.* (2016). 'Fair Data Technology Update'

About DCMI, https://www.dublincore.org/about/