Testing the FAIR metrics on data catalogs

“Metrics, Metrics can you recall,

Which data catalog is the FAIRest of them all?”

Jarno A A van Erp[1], Carolyn D Langen[1], Anca Boon[1], Kees van Bochove[1]

[1]The Hyve B.V., Arthur van Schendelstraat 650, 3511 MJ Utrecht

Corresponding Author:

Kees van Bochove[1]

Arthur van Schendelstraat 650, 3511 MJ Utrecht

Email address: office@thehyve.nl

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27151v2 | CC BY 4.0 Open Access | rec: 4 Sep 2018, publ: 4 Sep 2018

Abstract

The introduction of the FAIR –Findable, Accessible, Interoperable, Reusable– principles has caused quite an uproar within the scientific community. Principles which, if everyone adheres to them, could result in new, revolutionary ways of performing research and fulfill the promise of open science.

However, to bring about these changes, data users need to rethink the way they treat scientific data. Just passing a dataset along, without extensive metadata will not suffice anymore. Such new ways of executing research require a significantly different approach from the entire scientific community or, for that matter, anyone who wants to reap the benefits from going FAIR.

Yet, how do you initiate this behavioral change? One important solution is by changing the software scientists use and requiring data owners, or data stewards, to FAIRify their dataset.

Data catalogs are a great starting point for FAIRifying data as the software already intends to make data Findable and Accessible, while the metadata is Interoperable and relying on users to provide sufficient metadata to ensure Reusability. In this paper we analyse to what extent the FAIR principles are implemented in several data catalogs.

To determine how 'FAIR' a dataset is, the FAIR metrics were created by the GO-FAIR initiative. These metrics help determine to what extend data can be considered FAIR. However, the metrics were only recently developed, being first released at the end of 2017.

The Hyve has tested/evaluated three popular open source data catalogs based on the FAIR metrics: CKAN, Dataverse, and Invenio. Most data stewards will be familiar with at least one of these.

Within this white paper we provide answers to the following questions:

- Which of the three data catalogs performs best in making data FAIR?
- Which data catalog utilizes FAIR datasets the most?
- Which one creates the most FAIR metadata?
- Which catalog has the highest potential to increase its FAIRness, and how?
- Which data catalog facilitates the FAIRification process the best?

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27151v2 | CC BY 4.0 Open Access | rec: 4 Sep 2018, publ: 4 Sep 2018

Introduction

Earlier this year the international organisation GO-FAIR opened its first offices in Leiden, the Netherlands and Hamburg, Germany. The organisations aim is to promote and facilitate the implementation of the “internet of FAIR data and services”. Several countries, including Germany and the Netherlands, have committed themselves to implementing an infrastructure capable of supporting the newly arisen needs introduced by the FAIR data principles. This includes clearly defined data permissions (to ensure the research data is not misused), addition of extensive metadata, noninvasive data sharing, and eventually transmission and integration of data/information across different organisations in different countries with different laws and regulations. The efforts of GO-FAIR should result in a clear understanding across organisations what the requirements for each dataset are. Funding organisations, such as NWO and H2020, have already integrated requirements for FAIR data in their research grant application process, in this way safeguarding that data created with their funds will be reusable for future research.

One major concern regarding these new requirements is: How to facilitate making research data FAIR? FAIRifying data could become a time consuming activity when the FAIR data stewards have not sufficient insight in the research question or needs to revise a large amount of data. This increases the need for tools to identify FAIR business practices which require a low effort input while resulting in high value output. The tools that are most suitable for these low effort/high value FAIR business practices are data catalogs, as they are already used to make data findable and accessible, ensuring metadata is interoperable, and facilitating reuse of data.

Data catalogs

Combined with the vast amounts of data that need to be processed in scientific and medical research nowadays, one important system requirement is that data does not get lost. A common approach for preventing data loss is storing it in a data catalog. Data catalogs help to organize, structure and track metadata and data generated, so that the information can be saved and shared within an organisation. The use of data catalogs could even result in scientists getting more citations, as they create opportunities to elaborate on or reuse prior research. For example, a data catalog makes it much easier to search for relevant data.

To enable the entire scientific community to fully benefit from research data, the reusability of data should be improved in a trustworthy manner, protecting both the data producer and the external data re-user. By improving the quality and comparability of research data, fellow scientists should be able to reuse a particular dataset. Establishing trust between data producers and external data re-users is an issue requiring a stronger behavioural change among scientists than just asking them to add extra metadata to datasets. To facilitate the necessary change, several funding agencies nowadays require that grant recipients provide a Data Management Plan or Data Stewardship Plan, which includes a description of how the data will be made available to fellow researchers,,,.

Important aspects of data management with regard to the reusability of data are the location and method by which the data is stored. Besides, the risk of someone misusing sensitive data should be minimized and this aspect should be duly considered when making data available publicly or within an institute. Luckily, there are software solutions available that limit the risks of misuse, while ensuring the data can be reused by the original creator, the institute, and potentially anyone around the world. Ultimately, data reuse increases the value of datasets thus increasing the likelihood that (public) money is being spent in an efficient manner.

FAIR principles

Optimal data reusability is at the core of the FAIR principles. It ensures that humanity will get the most value out of publicly funded scientific research. In order to achieve this, the FAIR principles require metadata additions, data point annotations, and precise descriptions of potentially relevant information, thereby ensuring findability, accessibility, and safeguarding reusability by others. All these qualities combined make data catalogs the most attractive tools to start FAIRifying data.

However, their popularity means scientists can choose between a number of options. Many data catalogs have implemented changes to better adhere to the FAIR principles. It is rarely clearly stated, though, to which metrics they adhere and to what extend.

Background

To come up with a method of determining FAIRness of catalogs, we first need to define what exactly a data catalog is. Subsequently, we will discuss the method used to determine how FAIR the three catalogs are that The Hyve analysed.

Data catalogs

According to the Data Catalog vocabulary (DCAT), published by the World Wide Web Consortium (W3C), a data catalog is a curated collection of metadata about datasets. Practically speaking, this means a service which allows users to deposit, describe, share, discover and explore datasets. When data is correctly and accurately curated, they can be better understood and (re)used, which increases the value of any given dataset.

The Hyve evaluated three data catalogs using the FAIR metrics: CKAN, Dataverse and Invenio. The capabilities we explored include:

- Recognition of a variety of file formats
- Digital Object Identifier (DOI) generation
- Fast search indexing (e.g. using open source search engines SOLR or ElasticSearch)
- Harvesting (meta)data from external catalogs

The unique features of each catalog are described below.

CKAN

The CKAN (Comprehensive Knowledge Archive Network) catalog allows for the creation of "organizations" – entities which supply data. Organizations can contain multiple datasets, called assets within CKAN. An asset combines one or more files with metadata and tags. Views can be attached to the organisations allowing users to preview the assets using maps, graphs and tables (see Figure A). A news feed shows users recently added and/or modified assets.

CKAN has a plug-in system for adding features such as enhanced monitoring, custom user roles and permissions. At The Hyve we developed a plug-in expanding these options, focussing on a more fine grained accessibility mechanism within CKAN. The plug-in system adds custom metadata fields to datasets and allows users to search the data files. There is also a plug-in available that allows CKAN to expose and explore metadata from other catalogs using RDF documents serialized with DCAT and export CKAN data as RDF DCAT endpoints. This creates the option to register the catalog as a *FAIR data point*.

CKAN stores its metadata in a PostgreSQL database. Files uploaded to CKAN can be stored on a local drive, a network connected drive, or on Cloud storage solutions such as S3, Azure or Google Cloud Storage.

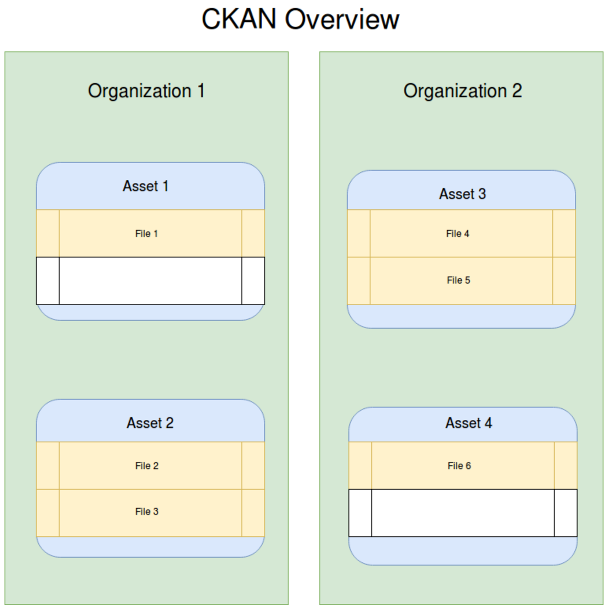


Figure A: CKAN structure. *CKAN allows for multiple organizations. Organizations consist of assets which contain one or more files.*

Dataverse

Dataverse works with data repositories which are called *dataverse* The catalog allows users to create a dataverse within a dataverse, where each dataverse has its own administration rights. As such, read and write permissions of each dataverse (and its datasets) can be controlled independently, and metadata can be assigned to a dataverse, a dataset, or to a single file. The recursive structure of a dataverse is illustrated in Figure E. Dataverse uses a PostgreSQL database, combined with Solr for searching, and a local file storage system for saving files.

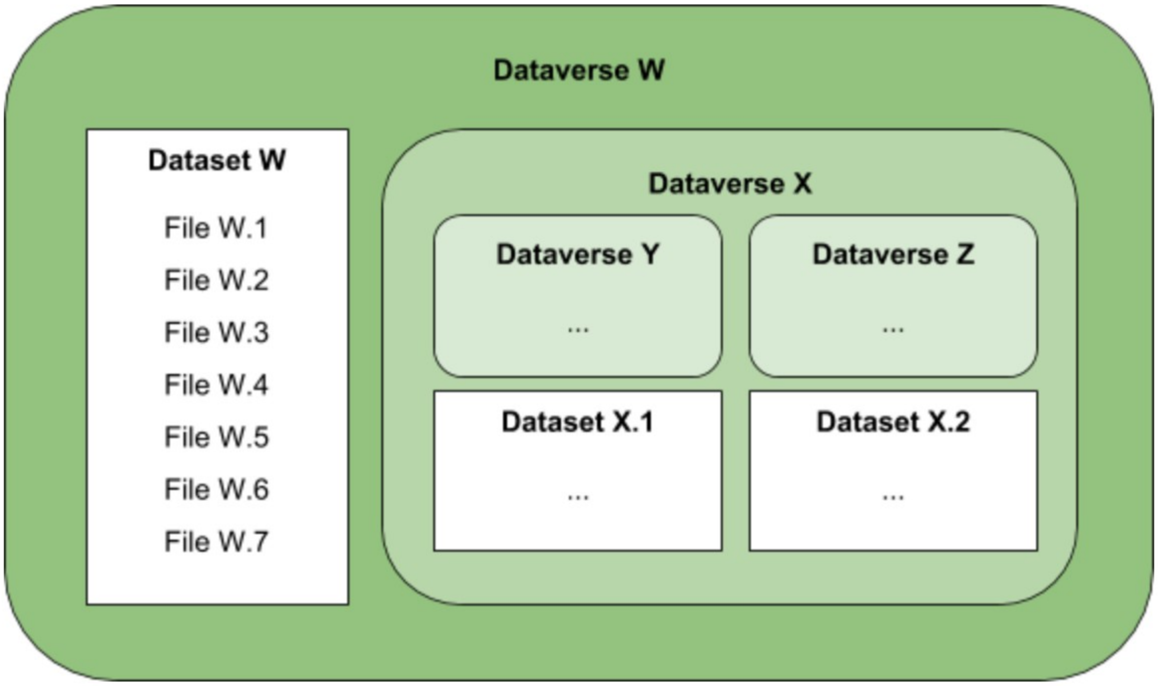


Figure B: Dataverse structure. *Dataverse is a data catalog containing datasets and possibly a number of Dataverses. Each dataset contains files. In this example, Dataverse W contains a dataset with seven files and Dataverse X, consisting of Dataset X.1 and X.2, Dataverse Y and Dataverse Z.*

Invenio

Invenio is a data catalog developed by CERN, the European Organization for Nuclear Research, to share their data publicly with fellow scientists. After more than fifteen years of experience with Invenio, CERN developers released a new, modularly-structured version, with three types of modules defined as base, core and additional feature. All modules are available in the Python Package Manager (PyPM) as separate components, which can be replaced by custom-made solutions.

The data model of Invenio (Figure X) consists of linking DOIs with a JSON (JavaScript Object Notation) Schema representation for the associated metadata. This grants a certain freedom to create links between datasets while at the same time limiting the complexity to a predefined model.

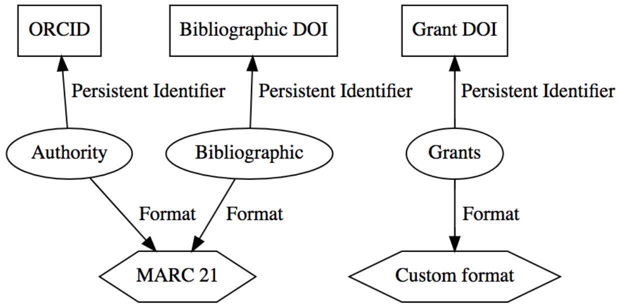


Figure C: Data model of Invenio. *Reproduced from: http://invenio.readthedocs.io/en/latest/developersguide/create-a-datamodel.html*

What is FAIR?

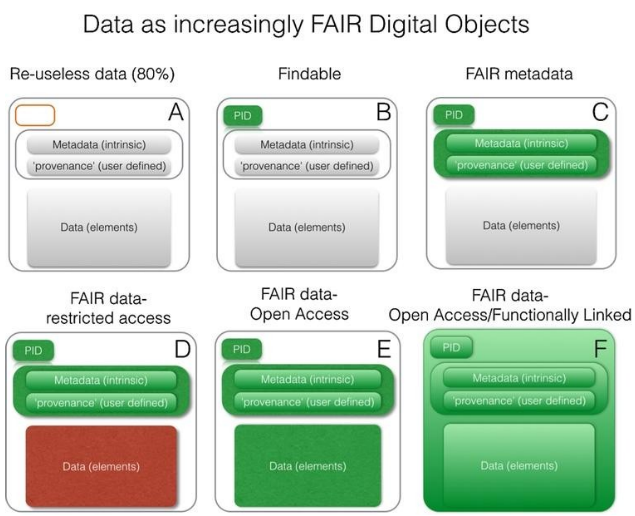
As mentioned before, making data FAIR –Findable, Accessible, Interoperable and Reusable- requires a behavioral change from scientists. Many researchers recognize that unwillingness to share results, loss of data and a focus on publication output instead of research quality are detrimental to the research community. By rating the data quality researchers produce rather than focussing on the ranking of the journal in which results are published, the GO-FAIR initiative aims to change the way scientific data is valued and wants to facilitate the process of behavioral change.

Ensuring that data can be reused creates a higher value proposition for generating data. By describing data with rich metadata and annotating the dataset itself, a computer, and eventually every machine, will be able to interpret the data: the machine "knows" what information the dataset contains, can link similar data, is able to create a knowledge graph out of these links. This would, all in all, reduce the time researchers spent searching for potentially interesting datasets. Besides, machines will be able to convert any created knowledge graph to human readable formats, enabling researchers to explore and use these more easily for research purposes. With the machine being able to identify what is inside any given dataset, it can make the data interoperable with other datasets. This leads to increased analytical capability and improved data maintenance.

To be able to assess the FAIRness of data, the FAIR metrics were developed. The metrics consist of multiple rules describing what is needed to comply with each one of the FAIR principles. For example, to be fully compliant to the first principle of Findability (F1) a Uniform Resource Locator (URL) to a registered identifier scheme should be provided, along with a URL that links to a document containing the relevant policy for when this identifier scheme becomes deprecated.

Degrees of FAIRness

The release of the FAIR principles created confusion about how to actually implement these standards. Degrees of FAIRness were introduced to bring clarity and explain if data needs to adhere to all criteria of a FAIR metric to even be considered FAIR. In this paper, we make suggestions how to use these gradients with regards to FAIR and end-user software (data catalogs in this case), contradicting a common perception that tools are either FAIR or not FAIR.



A data owner can determine to what extend he wants his data to be FAIR. For example, privacy sensitive patient data are never meant to be freely accessible. Figure D shows the guidelines that scientists should adhere to in order to obtain a certain degree of FAIRness in their datasets.

Figure D: Increasingly FAIR digital objects*. To be considered FAIR, some steps need to be taken. This figure gives an overview how different measures increase the FAIRness of a dataset.*

However, there is some discussion in the field if Figure F is a proper representation and description of the degrees of FAIRness, as a dataset for internal use can comply to only a number of FAIR principles and still be used in a FAIR manner. To make a dataset Findable to a degree, it could be sufficient to add just a PID (Persistent Identifier), metadata in a machine-readable format, and provenance.

Despite these shortcomings, the FAIR metrics do help determine the level of FAIRness of a certain dataset or data catalog. To truly be FAIR, all data and metadata should be stored in an online repository. Otherwise, how can someone else find and access them?

FAIR metrics

The FAIR metrics consist of four separate "groups", making the standard distinction between Findable, Accessible, Interoperable and Reusable, where each group has different metrics to determine if data adheres to the corresponding principle. This means that FAIR metrics FM-A1.1 corresponds to FAIR principle A1.1. This naming convention will be used in this paper, with the addition of FM referring to the metric related to the principle.

One aspect that needs to be emphasized, is that the FAIR metrics will be constantly changing and evolving with the introduction of new technologies and standards in years to come. Currently, the FAIR metrics have a strong emphasis on determining the FAIRness of a dataset rather than the FAIRness of software. For example, it is not stated where certain information needs to be located. In general, this additional information should be present within the dataset, although from a software perspective it should be enough to only link to certain, standard, information. If the metrics are used to determine the FAIRness of data contained within software, the way to find that data, or its location, needs to be stated clearly. When you want to assess the FAIRness of software and focus on automatic machine-readability, it currently is often unclear where the machines should search for specific information. Implementing this could lead to easier ways of connecting various FAIR tools.

As a tool to automatically determine the FAIRness of a data catalog is currently being developed by GO-FAIR, we performed our analyses by hand.

Data catalog FAIRness review

To determine the FAIRness of the three data catalogs, we looked at two ways data could be handled: manual and automated. Manual is defined as requiring additional effort from the reviewer, whereas automated meant that the data catalog can make machine readable exports from the data without the need for additional input from the user.

For both manual and automated scores we defined three outcomes per matrix: present, partial or absent. Present meaning that the catalog fully meets the criteria, partial meaning that it meets only part of the criteria, and absent implying it did not or not sufficiently meet the criteria. For each partial outcome, we identified what element was missing. An overview of missing FAIR elements of the three data catalogs can be found in Appendix A.

Leveraging the plugin design of CKAN, the ckanext-dcat plugin was added for automated DCAT exports. For Invenio the Zenodo version was used, as the FAIR metrics dataset was published in this version.

The overall outcome of the review is represented in Table A. The differences in findability, accessibility, interoperability and reusability between the three data catalogs are presented in Table Y. See Graph A for a visualization of Table Y.

	Dataverse		CKAN		Invenio		
Metric	Manual	Automated	Manual	Automated	Manual	Automated	Meaning automated
F1A	2	2	0	0	2	2	Automatically created
F1B	0	0	0	0	0	0	New IRI automatically findable when old identifier scheme becomes deprecated
F2	2	1	2	1	2	1	Metadata and data are automatically annotated in a machine-readable format
F3	2	2	2	2	2	2	IRI is automatically added to export of metadata
F4	2	2	2	2	2	2	When (meta)data is made public it is automatically indexed
A1.1	1	1	1	1	1	1	Protocol description is automatically added
A1.2	1	1	1	1	1	1	Machine knows if it does not have access and knows how to gain access and can perform the required action
A2	0	0	2	2	0	0	Metadata automatically stays when data is deleted
I1	2	1	2	1	2	1	Data is automatically annotated with a knowledge representation language
I2	2	1	2	1	2	1	Data is automatically annotated with FAIR vocabularies
I3	1	1	2	1	0	0	Relationships are automatically added to metadata, possible relationships are automatically discovered
R1.1	0	0	2	0	2	2	A machine automatically knows what it can and cannot do with the data
R1.2	2	2	2	0	2	2	Able to automatically create an overview of the provenance that meets the minimal requirements*
R1.3	0	0	0	0	0	0	Automated request for certification from a recognized body

Legend

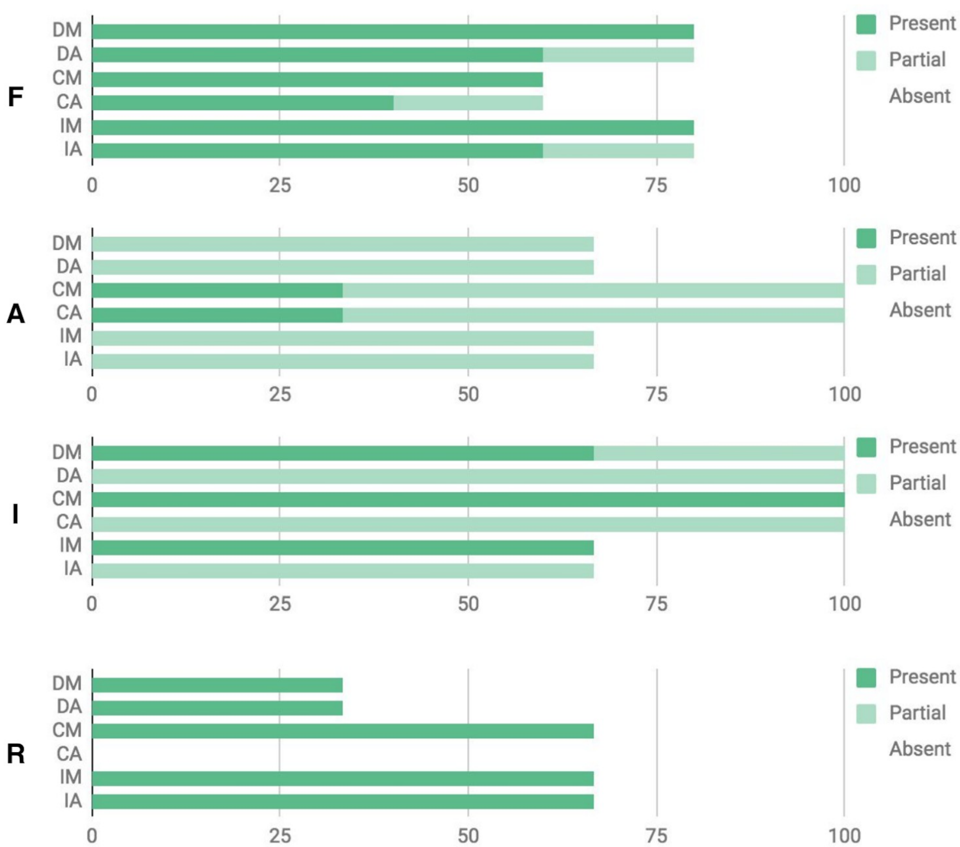
2	Present
1	Partial
0	Absent

*R1.2 minimal requirements:
- Who/what/When produced the data (i.e. for citation)
- Why/How was the data produced (i.e. to understand context and relevance of the data)

Table A: *Outcome of the FAIR metrics review per metric. The row on the right side specifies what was expected from automated data handling.*

	Dataverse		CKAN		Invenio	
	Manual	Automated	Manual	Automated	Manual	Automated
Present:	7	4	9	3	8	5
Partial:	3	6	2	6	2	5
Absent:	4	4	3	5	4	4
Total (max=14):	**8.5**	**7.0**	**10.0**	**6.0**	**9.0**	**7.5**

Table B:. *General outcome of the FAIR metrics review for data catalogs. A total of 14 points could be granted to each tool. Total was calculated as follows: Total = present+(partial/2)*



Graph G: *Visual representation of the FAIR metrics review outcome. See Table A for details.*

Thanks to the liberty CKAN offers, it can be judged the FAIRest of the three catalogs. However, depending on the choices users make, there will be significant differences in the FAIRness of data. A downside of using CKAN (originally an open data repository) is that it offers by default no authentication mechanism, which is an important functionality for scientific communities (especially those dealing with data from human subjects) with laws such as the General Data Protection Regulation (GDPR). Yet despite this shortcoming, CKAN does have the highest score for manual FAIRness.

Regarding the FAIRness of Dataverse and Invenio, the major differences are that Invenio provides better permission support. As a downside it does not link datasets adequately. Dataverse includes a license, but it only specifies the licence name. This provides insufficient information. Dataverse does allow the user to create a link between datasets, but this feature can be considered too limited as it provides not enough options to describe meaningful relationships.

Discussion

When we combine the results for both manual and automated generation of FAIR data, CKAN performs worse than both Dataverse and Invenio. Therefore, we conclude that CKAN is good at handling data that is already FAIR. However, the catalog will be less helpful in the data FAIRification process. This draws our attention to different angles that software FAIRness can be viewed from and raises the question if a data catalog is more FAIR when it helps users in making their data FAIR or when it supports data that is already FAIR, without facilitating of the FAIRification process?

The FAIR metrics

One remarkable outcome of our FAIR review is the low score on the metrics evaluating accessibility, with Dataverse and Invenio both performing worse than CKAN. When looking at the specific criteria the data catalogs did not meet, it becomes clear that CKAN only scores higher because of its ability to preserve metadata when the original dataset has been removed (principle A2). As for the other two metrics, FM-A1.1 and FM-A1.2, all three data catalogs achieve only a partial score. When all data catalogs render the same partial score, it raises the question if this is a result of the FAIR metrics not being defined clear enough or because the proposed methods in the FAIR metrics have not yet been implemented.

The question of unclear definition versus incomplete implementation comes to mind especially when assessing the FAIRness of software, as the FAIR principles are not only meant to be guidelines, but also set out a roadmap to what should eventually be implemented. For example, FAIR principle R1.3 states: "(meta)data meet domain-relevant community standards". The community around data catalogs is the scientific publishing and data preservation community, yet it is currently doubtful if this is the same community that determines what the domain-relevant standards are. Even within a specific field such as the life sciences, there are many standard sets for metadata to choose from (e.g. schema.org, specifications such as DATS and DCAT or the HCLS Community Profile). The associated metric (FM-R1.3) also states that "a valid result" is given, when there is a "successful signature validation". The result is based upon "a certification saying that the resource is compliant", which needs to be provided. But at the moment it is unclear who is authorized to give this certification and/or signature. Can someone just make a certificate and say "This is my community, therefore I set the standards and decide that I meet those criteria"? Or is there need for an external body to certify compliance? Should the communities establish the standards and certification bodies themselves with the help of GO-FAIR?

Currently, it seems the FAIR metrics outline a prevalent struggle within the FAIR community: inclusiveness or interoperability through standardization? Using standards for interoperability automatically means everyone not using those standards will be excluded. A solution could be to map these standards, an effort already started by FAIRsharing. However, this current mapping project could very well never be finished with new standards emerging that might be more widely adopted. This could be an everlasting discussion. Therefore, we will use what is currently available for reviewing the FAIRness of software and leave this discussion to the FAIR community.

Data catalogs and FAIRness

Findability

The main function of data catalogs is providing an overview of the available data. It is therefore not surprising that all data catalogs score high on the Findability metric. What they all are lacking, though, is a procedure for when the original identifier scheme becomes deprecated. The catalogs also lost half a point for "the availability of machine-readable metadata that describe a digital resource" with regard to automatic creation of FAIR data in certain formats. At this moment, metadata still needs to be added manually by the user.

Accessibility

The Accessibility metrics has a surprisingly low score. This can be a result of the way the metrics have been formulated or because reality has not yet caught up with the FAIR metrics, depending on which metric is reviewed. Accessibility, for example, has three subcategories: FM-A1.1,FM-A1.2 and FM-A2. While the latter is clear and makes sense: "The existence of metadata even in the absence/removal of data", the right way to implement the first one is unclear: "The nature and use limitations of the access protocol."

If the secure communication protocol HTTPS is used for data transfer, the metadata export should include a description of the HTTPS protocol as per FM-A1.1. However, if the machine is already familiar with the protocol, does a description still need to be included? FM-A1.2, "Specification of a protocol to access restricted content", is an example of a metric that has not yet been implemented in any data catalog. Implementation of this metric would require a machine that automatically knows how to access a data source and execute the task needed to gain access.

Interoperability

For interoperability, all data catalogs reach partial to high scores, depending on whether the automatic or manual score is consulted. This outcome is as expected. Fully automated interoperability would mean automatic detection of what the dataset contains and linking this to an already existing knowledge base. This information would then be used to automatically create links to vocabularies and ontologies and search for relations between datasets. To be able to perform such a complicated task, which needs to be accurate as well, some serious AI power is needed. This, in it's turn, requires a large number of FAIR datasets to train the AI. Unfortunately, neither of these two requirements are currently met. Therefore, asking the question "How do data catalogs perform on automated interoperability?" is at the moment nonsensical. For now, a better question would be "Does this data catalog link to a machine-readable version upon addition of a dataset?", when interoperability is a necessity or high priority.

Reusability

When it comes to reusability the data catalogs all achieve a low score. This finding is disturbing, as one purpose of data catalogs is to ensure the reuse of data. As discussed above, FM-R1.3 is formulated in such a way that full reusability cannot currently be achieved. The other two reusability metrics emphasize licensing and provenance. Although licenses can be added to all three catalogs, within CKAN adding a license is optional. The same goes for provenance. In CKAN the user has the option to specify information. Since this should be a requirement for FAIR data, CKAN achieves a lower score for automated data FAIRification.

Manual versus automated

All three evaluated catalogs are at least in part capable of handling user-created FAIR data. Manually making a dataset adhere to the the highest FAIRness standards requires a significant effort. A data catalog which supports FAIR data without automation is therefore not a serious contender when deciding which data catalog best meets the user's needs. This is an area where these catalogs can improve upon significantly.

Another factor to consider when determining which data catalog to use, is to what extent the tool helps the user to make data FAIR. One of the easiest ways to achieve this is by adding configurable fields, allowing an administrator to decide which fields are mandatory and which are optional. This ensures that the necessary metadata are added, allowing the tool to automatically convert the information into a machine-readable format.

Conclusion

Although CKAN with manual FAIRification of data by the user has the highest single FAIR metrics score, it does not score highest overall. This is because the catalog does not help and guide users to make data FAIR as much as the other tools. A catalog should help the user with making their data FAIR by default as much as possible.

This is where the historical differences between Invenio and Dataverse become apparent. Dataverse was created as a data catalog for researchers, whereas Invenio was developed for storage of bulk data. This results in Dataverse focusing on storing data in such a way that it can be used for publication, whereas Invenio provides a higher quality implementation of the FAIR data principles that ensure trustworthy reusability.

Taking everything into account, our overall conclusion is that CKAN can handle data that is already made FAIR better. While Dataverse can be considered just as FAIR as Invenio. However, the ultimate goal of the FAIR initiative is, as mentioned, to change the behavior of researchers and data stewards handling data and have them reconsider how to publish the data they create: ensuring high quality metadata are added and establishing trust regarding reusability by clearly defining rules and developing guidelines for access and permissions. With respect to these aspects, Dataverse is the one that outshines the other two catalogs.

Next steps?

The low score for Accessibility surprised us. However, bear in mind that score resulted from a number of FAIR requirements that cannot be fully met at the moment. What can we learn from this? In the first place, the low score indicates that the techniques used for granting and retracting access have not yet been fully developed, implemented and accepted for usage according to the FAIR data principles. This does not mean that this aspect is overlooked. As a software developer, you for example may assume that “machine readable accessibility” means that servers and/or clients are automatically authorized and authenticated based upon the identity provided, together with the request made.

The high overall scores for Findability reassures that these principles are already widely accepted and implemented. This means that the scientific community is ready for the next step: accessibility. Ideally, this would mean creating an infrastructure where (external) researchers can request data and only need to accompany this request with a link to a verified online identity. A number of scientists have already expressed an interest in supporting external access and identity providers, with techniques such as OAuth and websites like ORCID.

Manual vs Automated or Reality vs Future?

In this paper, two types of FAIRness were discussed with regards to software: manual and automated. However, this raises the question if it is even realistic to demand fully automated FAIR data creation. Of course, such techniques are not available yet and might not be for some time to come.

Therefore, a more sensible question would be: Do we expect researchers, for the time being, to manually add all the metadata needed to make their data FAIR? Can this be a task be left to a machine? Imagine AI being able to define and add FAIR metadata based on headers, column names, data published in research papers, drafts of scientific papers, et cetera. This might not result in full FAIRness but it would lift the burden currently resting on the researchers' shoulders. For such a task AI needs to be created and trained, and for this training FAIR datasets are needed. This raises the issue that such datasets are simply not available right now in sufficient numbers.

As mentioned, data catalogs are among the best tools to implement the FAIR data principles. In order to exploit them to the fullest, it is necessary to standardize the requirements, ensuring that data stored across catalogs is universally findable and accessible. To eventually ensure that data in a data catalog is findable and obtainable without the user actually seeing the user interface of the data catalog. By making the addition of certain (meta)data mandatory, data catalogs can play a vital role in the creation of FAIR metadata sets. With these datasets an AI specialized in tagging data with FAIR metadata can be created and trained. Eventually, with a large enough number of FAIR datasets, this AI would be able to annotate every dataset in a FAIR way, completely lifting this burden off the researchers' shoulders.