Alignment of biomedical data repositories with open, FAIR, citable and trustworthy principles

Fiona Murphy[1]*, Michael Bar-Sinai[2]*, Maryann E. Martone[3]

[1]MoreBrains Cooperative Ltd, Chichester, UK
[2]Department of Computer Science, Ben-Gurion University of the Negev and The Institute of Quantitative Social Science at Harvard University
[3]Department of Neurosciences, University of California, San Diego; SciCrunch, Inc.

*Contributed equally to this manuscript

Fiona Murphy*: https://orcid.org/0000-0003-1693-1240
Michael Bar-Sinai*https://orcid.org/0000-0002-0153-8465
Maryann E. Martone[3]: https://orcid.org/0000-0002-8406-3871

Abstract

Increasing attention is being paid to the operation of biomedical data repositories in light of efforts to improve how scientific data is handled and made available for the long term. Simultaneously, groups around the world have been coming together to formalize principles that govern different aspects of open science and data sharing.The most well known are the FAIR data principles. These are joined by principles and practices that govern openness, citation, credit and good stewardship (trustworthiness). Together, these define a framework for data repositories to support Open, FAIR, Citable and Trustworthy (OFCT) data. Here we developed an instrument using the open source PolicyModels toolkit that attempts to operationalize key aspects of OFCT principles and applied the instrument to eight biomedical community repositories listed by the NIDDK Information Network (dkNET.org). The evaluation was performed through inspection of documentation and interaction with the sites. Overall, there was little explicit acknowledgement of any of the OFCT principles, although the majority of repositories provided at least some support for their tenets.

Introduction

Best practices emerging from the open science movement emphasize that for data to be effectively shared, they are to be treated as works of scholarship that can be reliably found, accessed, reused and credited. To achieve these functions, the open science movement has recommended that researchers formally publish their data by submitting them to a data repository (OpenAire 2020), which assumes stewardship of the data and ensures that data are made FAIR: Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016). Pub-

lishing data can therefore be seen as equivalent to publishing narrative works in that the locus of responsibility for stewardship transfers from the researcher to other entities, who ensure consistent metadata, future-friendly formats, stable and reliable access, long term availability, indexing and tools for crediting the contributors. As these types of responsibilities are traditionally supported by journals and libraries, it is not surprising that many publishers and libraries are now developing platforms for hosting research data. At the same time, data are not exactly the same as narrative works. They require additional functionality to increase their utility, which explains why the most well known scientific data repositories are led by individual researchers, research communities or funders. Scientific data repositories such as the Protein Data Bank (Berman et al. 2012) predated the internet and are viewed as important infrastructures for data harmonization, integration and computation.

Although there is general agreement that repositories should support FAIR data, there have been several other community-led initiatives to develop principles in support of open science and data sharing. The "Defining the Scholarly Commons" project at FORCE 11.org identified over 100 sets of principles issued by organizations and groups around the world that cover a range of activities involved in scholarship and how it should be conducted in the 21st century (Bosman et al., 2017). Common threads included: 1) the need to include not only narrative works, but data, code and workflows; 2) the desire to make these products "as open as possible; as closed as necessary"; 3) FAIRness, i.e., designing the products of scholarship so that they operate efficiently in a digital medium; 4) Citability, i.e., expanding our current citation systems to cover other research outputs like data, and 5) Trustworthiness, i.e., ensuring that those who assume responsibility for stewardship of scholarly output operate in the best interests of scholarship. In the imagined scholarly commons, data repositories were the central players that provided the human and technical infrastructure for making research data Open, FAIR, Citable and Trustworthy (OFCT).

In the work presented here, we developed an instrument to assess the current state of data repositories on behalf of the NIDDK Information Network (dkNET.org; (Whetzel et al. 2015)). dkNET was established in 2012 to provide information and services to basic and clinical bio-medical researchers for data and resources relevant to diabetes, digestive and kidney diseases (referred to here as "dk"). dkNET is taking an active role in interpreting and facilitating compliance with FAIR on behalf of this community. Part of this effort involves creating tools to help researchers select an appropriate repository for their data. As a first step, dkNET created a listing of data repositories that cover domains relevant to dk science as listed on dkNET's own website. As a second step, we wanted to evaluate how well these repositories supported current trends in open science. We therefore developed an instrument that allowed us to gauge repositories' alignment with OFCT principles.

Principle	Description	Guiding principles/charters
Open	Research outputs should be as open as possible and as closed as necessary	Open Definition 2.1(Open Knowledge Open Definition Group 2020)
FAIR	Research outputs should be designed to be Findable, Accessible, Interoperable and Reusable for humans and computers	FAIR Data Principles (Wilkinson et al. 2016)
Citable	Research outputs should be supported by formal systems of citation for the purposes of provenance and credit.	Joint Declaration of Data Citation Principles (JDDCP) (Data Citation Synthesis Group 2013); Software Citation Principles
Trustworthy	Data repositories should demonstrate that they are responsible for long term sustainability and access of data entrusted to them	Principles of Open Infrastructures (Bilder, Lin, and Neylon 2015); Core Trust Seal (CoreTrustSeal Standards and Certification Board 2019)

Table 1: Guiding principles for OFCT used in this study to develop the assessment instrument

Method

We developed a set of 31 questions (Table 2) operationalizing the major elements of each of the principles listed in Table 1. We did not attempt to cover all aspects of the principles, but selected those that were relevant for repositories and for which clear criteria could be developed. At the time we conducted this study, the TRUST principles had not yet been issued and so are not included explicitly in our instrument, although much of what is covered in the CoreTrustSeal is relevant to the TRUST principles.

Q#, id	*Question text*	*Answers*	C	D	P
1 acc	**Does the repository provide access to the data with minimal or no restrictions?**	no restrictions minimal restrictions significant restrictions significant but not justified	N		O
2 reuse	**Are you free to reuse the data with no or minimal restrictions?**	yes somewhat no	N		O
3 lic-clr	**Does the repository provide a clear license for reuse of the data?**	dataset level repository level no license	N		F
4 lic-cc	**Are the data covered by a commons-compliant license?**	best good somewhat open closed	Y	**lic-clr**	O
5 plat	**Does the repository platform make it easy to work with (e.g. download/re-use) the data?**	yes no	N		F
6 ru-doc	**Does the repository require or support documentation that aids in proper (re)-use of the data?**	best good adequate lacking	N		F
7 sch-ui	**Does the repository provide a search facility for the data and metadata?**	yes no	N		F
8 pid-g	**Does the repository assign globally unique and persistent identifiers (PIDs)?**	yes no	N		F

9 orcid	**Does the repository allow you to associate your ORCID ID with a dataset?**	required supported not available	N		C
10 md-level	**Does the repository support the addition of rich metadata to promote search and reuse of data?**	rich limited minimal	N		F
11 md-prv	**Are the (meta)data associated with detailed provenance?**	best good worst	N		F
12 md-daci	**Does the repository provide the required meta-data for supporting data citation?**	full partial none	N		C
13 md-ref	**Do the metadata include qualified references to other (meta)data?**	best good worst	N		F
14 md-lnk	**Does the repository support bidirectional link-ages between related objects such that a user accessing one object would know that there is a relationship to another object?**	best good unclear worst	N		F
15 fmt-com	**Does the repository enforce or allow the use of community standards for data format or meta-data?**	yes no	N		F
16 md-dkn	**Does the repository accept metadata that is applicable to the dkNET community disciplines?**	best good worst	N		F
17 md-psst	**Does the repository have a policy that ensures the metadata (landing page) will persist even if the data are no longer available?**	no by evidence by policy	N		F
18 md-FAIR	**Do the metadata use vocabularies that follow FAIR principles?**	enforced allowed minimal	N		F
19 land-ctsp	**Does the machine-readable landing page support data citation?**	yes no	N		C
20 md-cs	**Does the repository use a recognized community standard for representing basic metadata?**	yes no	N		F
21 acc-api	**Can the (meta)data be accessed via a standards compliant API?**	yes no	N		F

22 md-vcb	**Do the metadata use a formal accessible shared and broadly applicable language for knowledge representation?**	yes no	N		F
23 sch-api	**Does the repository provide API-based search of the data and metadata?**	yes no	N		F
24 gov-tsp	**Is the governance of the repository transparent?**	best good worst	N		T
25 oss	**Is the code that runs the data infrastructure covered under an open source license?**	best good no	N		T
26 tr-seal	Has the repository been certified by Data Seal of Approval or the Core Trust Seal or equivalent?	yes no	N		T
27 gov-stk	**Is the repository stakeholder governed?**	full good weak none	N		T
28 land-api	**Does the repository provide a machine-readable landing page?**	yes no	Y	**land-pg**	F
29 land-pg	**Does the PID or other dataset identifier resolve to a landing page that describes the data?**	yes no	Y	**pid-l**	C
30 md-pid	**Does the metadata clearly and explicitly include identifiers of the data it describes?**	all some none	Y	**land-pg, pid-l**	F
31 pid-l	**Does the repository assign, or the contributor provides, a locally unique identifier to the dataset or the data contribution?**	yes no	Y		F

Table 2: Questions and properties used for the final interview, The table shows the question order and ID (Q#,id), the text of the question posed in the interview (Question text), possible answers (Answers), whether or not the question is conditional ("C"), the dependencies of conditional questions (D) and the principle(s) the question is meant to cover (P). A "Y" in the conditional column indicates that whether or not the question is shown to the interviewer depends upon a prior answer. The questions that elicit the conditional questions are shown in the Dependencies column. Y=Yes, N=No, O=Open, F=FAIR, C=Citable, T=Trustworthy. The full instrument, which also includes explanatory text and appropriate links, is available at Martone et al., 2020.

The instrument was used to evaluate eight repositories listed by dkNET (RRID:SCR_001606) provided in Table 3. We selected these repositories to represent different data types or different research foci. Excluded from consideration were repositories that required an approved account to access the data, e.g., the NIDDK Central Repositories. We also did not con-

sider knowledge bases, defined here as a database that extracts observations from the literature or as a result of analyses of primary data, but not the primary data themselves. We also excluded some of the most well known of the biomedical databases, e.g., the Protein Data Bank and GEO, in order to focus on more dk-specific repositories. We included two generalist repositories, Zenodo and NIH-Figshare, as the generalist repositories are likely to play an increasingly significant role for diverse domains like dk, where specialist repositories for all data types and research foci may not be available. NIH-Figshare at the time of evaluation was made available as a pilot by the National Library of Medicine for data deposition by NIH-supported researchers, Many of these repositories are complex websites with multiple tools, services and databases, and so for each of the repositories, we indicate in Table 3 which specific component(s) were reviewed.

Repository	RRID	Description	Section	URL
AMP-T2D (Accelerating Medicines Partnership - Type 2 Diabetes Knowledge Portal)	RRID:SCR_003743	Portal and database of DNA sequence, functional and epigenomic information, and clinical data from studies on type 2 diabetes and analytic tools to analyze these data.	Data	http://www.kp4cd.org/datasets/t2d
Cell Image Library	RRID:SCR_003510	Freely accessible, public repository of vetted and annotated microscopic images, videos, and animations of cells from a variety of organisms, showcasing cell architecture, intracellular functionalities, and both normal and abnormal processes.	Main site representing single image and datasets	http://www.cellimagelibrary.org

Flow Repository	RRID:SCR_013779	A database of flow cytometry experiments where users can query and download data collected and annotated according to the MIFlowCyt data standard.	Public site	http://flowrepository.org
Image Data Resource (IDR)	RRID:SCR_017421	Public repository of reference image datasets from published scientific studies. IDR enables access, search and analysis of these highly annotated datasets.	Cell-IDR	http://idr.openmicroscopy.org/cell/
Mass Spectrometry Interactive Virtual Environment (MassIVE)	RRID:SCR_013665	MassIVE is a community resource developed by the NIH-funded Center for Computational Mass Spectrometry to promote the global, free exchange of mass spectrometry data.	Access public datasets	https://massive.ucsd.edu/ProteoSAFe/datasets.jsp#%7B%22query%22%3A%7B%7D%2C%22table_sort_history%22%3A%22createdMillis_dsc%22%7D
Metabolomics Workbench	RRID:SCR_013794	Repository for metabolomics data and metadata which provides analysis tools and access to various resources. NIH grantees may upload data and general users can search metabolomics database.	Data Repository	https://www.metabolomicsworkbench.org/data
NIH Figshare	RRID:SCR_017580	Repository to make datasets resulting from NIH funded research more accessible, citable, shareable, and discoverable.	Public portal and password protected space	https://nih.figshare.com/
Zenodo	RRID:SCR_004129	Repository for all research outputs from across all fields of science in any file format.	Public site and data submission forms	https://zenodo.org/

Table 3: List of repositories evaluated in this study. The specific section of the repository evaluated is indicated in the Section column, along with the corresponding URL.

Developing and testing the instrument:

To design the instrument, we adapted the decision tree originally designed by the FORCE11 Scholarly Commons project for evaluating repositories on OFCT principles (Bosman et al., 2017). We benchmarked the instrument against a range of surveys and other tools then available for similar uses. These included the repository finder tool developed by DataCite for the Enabling FAIR Data project; the Scientific Data journal repository questionnaire; the FAIRsFAIR data assessment tool; and the Core Trustworthy Data Requirements. From this exercise, we determined that the answers to the questions were sometimes difficult to ascertain as clear criteria for evaluation had not been specified. Some areas were clearly missing while some of the questions were duplicative. We thus modified the questionnaire by removing duplicates, adding additional questions, developing specific evaluation criteria and adding tips as to where to look for certain types of information. Definitions and links to supporting materials were also provided for each question where appropriate. The complete version of the questionnaire used here, which includes the criteria used for each question, was deposited in Zenodo (Martone, Murphy, and Bar-Sinai 2020)

The final questionnaire comprised 31 questions, listed in order in Table 2. Some of the questions are conditional, that is, their presentation is dependent upon a prior answer. For example, if an interviewer answered "No" to question **lic-clr,** "Does the repository provide a clear license for reuse of the data?" then question **lic-cc** "Are the data covered by a commons-compliant [i.e., open] license?" is not presented. Thus, the total number of questions asked may differ across repositories.

Table 2 also lists the principle set it covers (OFCT). Although the questions were originally grouped by principle, when testing the questionnaire we noted that many questions were logically related to one another, e.g., under the FAIR section we asked about licenses, while under the open section we asked about open licenses. Therefore, we reordered the questions to reflect better the actual workflow a reviewer might implement by grouping together related questions.

Encoding the instrument in policy models: The questionnaire was encoded using the PolicyModels software (RRID:SCR_019084). PolicyModels uses formal modeling to help humans interactively assess artifacts or situations against a set of rules. A PolicyModels model consists of an n-dimensional space (called "policy space"), and a decision graph that guides users through that space using questions. Each of the policy space's dimensions describes a single assessed aspect using ordinal values. Thus, every location in a policy space describes a single, discrete situation with regards to the modeled guidelines (M. Bar-Sinai, Sweeney, and Crosas 2016).

The dimensions of the policy space defined for this work formally capture the assessment aspects implied by OFCT. It contains 45 dimensions that are assessed by the 31 questions shown in Table 2, such as Documentation Level (lacking/adequate/good/full), Metadata Provenance (unclear/adequate/full), and overall ratings of each criteria, e.g., FAIR Accessibility level (none/partial/full) and so forth. The full policy space for this instrument is shown in Figure 1, and is also available via the questionnaire landing page and in Martone et al., (2020). Some dimensions are assigned based on the answer to a single question, while some are calculated based on values on other dimensions. Using an interactive interview guided by our model's decision graph, we were able to find the location of each of the evaluated repositories in the space we defined. To visualize this space, we developed an interactive viewer available at

http://mbarsinai.com/viz/dknet. This allowed us to formally compare repositories across multiple dimensions, and to collect overall statistics.

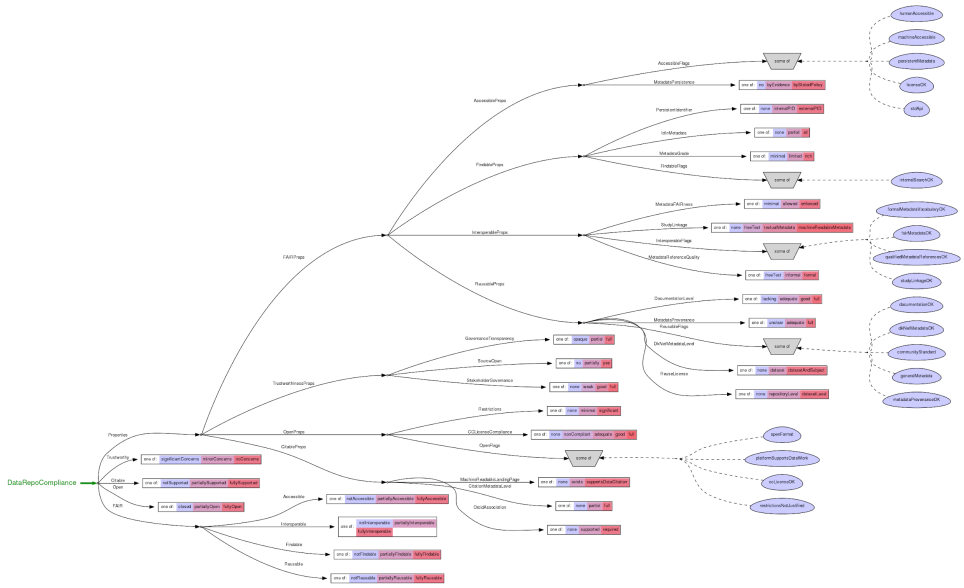

Figure 1: Policy space defined by the PolicyModels software illustrating the relationship of the dimensions assessed to the properties (rectangles) and flags (blue ovals). A full resolution view is available in Martone et al., (2020).

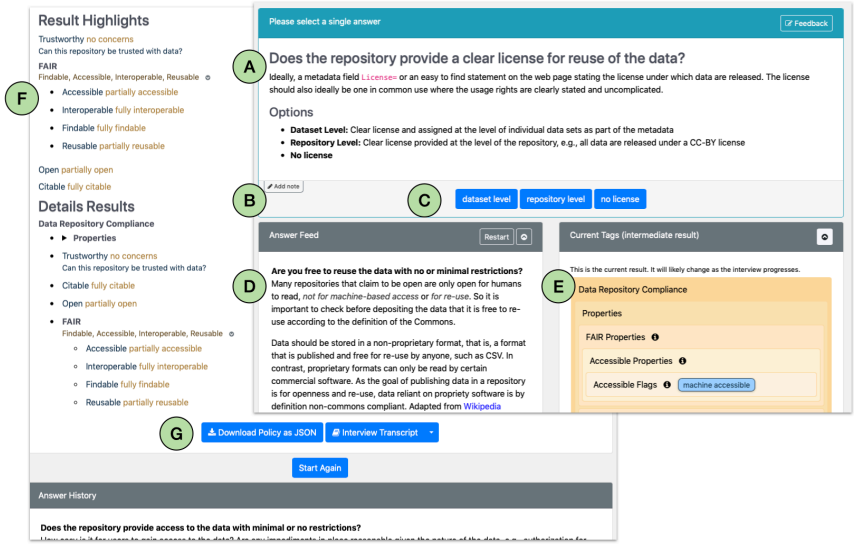

Figure 2: Main features of Policy Models questionnaire. The panel on the right provides an example of the question panel and the left panel shows the results of a survey after it is completed. A) each question is presented in sequence and can be accompanied by explanatory material and links to additional material; B) The interviewer may add notes to each question; C) Interviewer records an answer by selecting the appropriate response; D) The answer feed may be displayed and used to track progress and also to allow an interviewer to revisit a question to change an answer; E) Policy models tallies the answers and assigns tags assessing compliance with OFCT; F) Final tags assigned for each category; G) The results may be downloaded as json or xml.

The main features of the tool are shown in FIgure 2. The online version allows interviewers to annotate the response to each question with notes (Figure 2B) and export the outcomes of the evaluation (Figure 2G). Currently, the results can only be exported as .json or .xml. However, to save a human readable version .pdf version of the questionnaire results, users can use the browser's print function to save the interview summary page as a PDF.

Scoring

Five of the sites were reviewed independently by FM and MM between March and May 2020 and three in December 2020. Results were compared and a final score assigned for each question. The reviewers made a good faith effort to find information on the site to provide an accurate answer for each question. The evaluation included checking of information on the

repository site, examination of the metadata provided by the site, investigations into the PID system, including what information was exported to DataCite if DOIs were used, inspection of the underlying platform code, documentation and tutorials. For some of the repositories, we created accounts in order to evaluate practices and further documentation for uploading data, e.g, can one associate an ORCID with a dataset, although in no case did we actually upload any data. We did not attempt to read papers that described the site. If we could not find explicit evidence for a criterion, we assumed that it was not present. Therefore, a “No” answer to a question such as “Does the repository provide an API” could mean either that the repository has a statement saying that it will not provide an API, or that we could find no evidence that it did.

After a model-based interview regarding a given repository is completed, PolicyModels displays a coded evaluation of the repository. Formally, PolicyModels locates the coordinate that best describes that repository in our model’s policy space. While mathematically all dimensions are equally important, PolicyModels allows its users to organize them hierarchically, to make working with them more comfortable.

Our proposed model’s policy space is organized as follows. High-level property descriptions, such as openness and citability levels, are each represented in a dimension of their own. These dimensions have three levels, corresponding to “not at all”, “somewhat”, and “fully”. For example, the Reusable dimension contains the levels “not reusable”, “partially reusable”, and “fully reusable”.

The high-level properties are a summary of lower-level assertions, each describing a narrow aspect of these high-level properties. These assertions can be binary or detailed. For example, “open format”, one of the openness sub-aspects, is “yes” for repositories that use an open format and "no" for the others. On the other hand, “Study Linkage”, an interoperability

sub-aspect, can be “none”, “free text”, “textual metadata”, or “machine readable metadata”.

Each interview starts by pessimistically setting all high-level dimensions to their lowest possible value: "not at all". During the interview, while lower-level aspect results are collected, high-level repository coordinates may be advanced to their corresponding "somewhat" levels. After the last question, if the evaluated repository achieved an acceptable for all sub-aspects of a certain higher property, that property is advanced to its "fully" level.

As a concrete example, consider the "Findable" dimension. At the interview’s start, we set it to "not findable". During the interview, our model collects results about persistent identifiers used by the repository (none/internal/external), the grade of the metadata it uses (minimal/limited/rich), whether ids are stored in the metadata (none/partial/all), and whether the repository offers an internal search feature (yes/no). If a repository achieves the lowest values in all these dimensions, it maintains its "not findable" score. If it achieves at least one non-lowest value, it is advanced to "partially findable". After the interview is completed, if it achieved the highest value in each of these dimensions, it is advanced to "fully findable".

Data and Code Availability

The data outputs and completed questionnaires from the interview analysis are in Zenodo (RRID:SCR_004129): https://zenodo.org/record/4069364.

The latest version of the dkNET evaluation instrument is available at http://trees.scicrunch.io/models/dkNET-DRP/start and is made available under a CC-BY 4.0 license. The version used for this study, V1.0 is available at: http://trees.scicrunch.io/models/dkNET-DRP/7/?localizationName=en-US. A copy of the codebook for the instrument along with the visual-

izations produced by the PolicyModels software is available through Zenodo at (Martone, Murphy, and Bar-Sinai 2020)

Additional explication of the Policy Models dimension usage: https://github.com/codeworth-gh/dkNET-DecisionTrees/blob/master/data-repo-compliance/dimension-usage.adoc

A snapshot of the code underlying this study is available at: https://doi.org/10.5281/zenodo.4275004

PolicyModels is managed in GitHub (https://github.com/IQSS/DataTaggingLibrary) under an Apache v2 Open-source license. The summary tools (https://github.com/michbarsinai/PolicyModelsSummary) are released under an MIT license.

Results

Overall impressions

Figure 3 provides the average score, scaled to a 10 point scale for each question, with 1 = lowest score and 10 = best score. A full list of question IDs is available in Table 3 and Supplemental Material S1. On over half of the questions (17/31), repositories scored on average higher than the midpoint, indicating at least some alignment. On just under half they were below (14/31), indicating poor alignment or no information available, with all repositories receiving the lowest score on 3 of the questions.

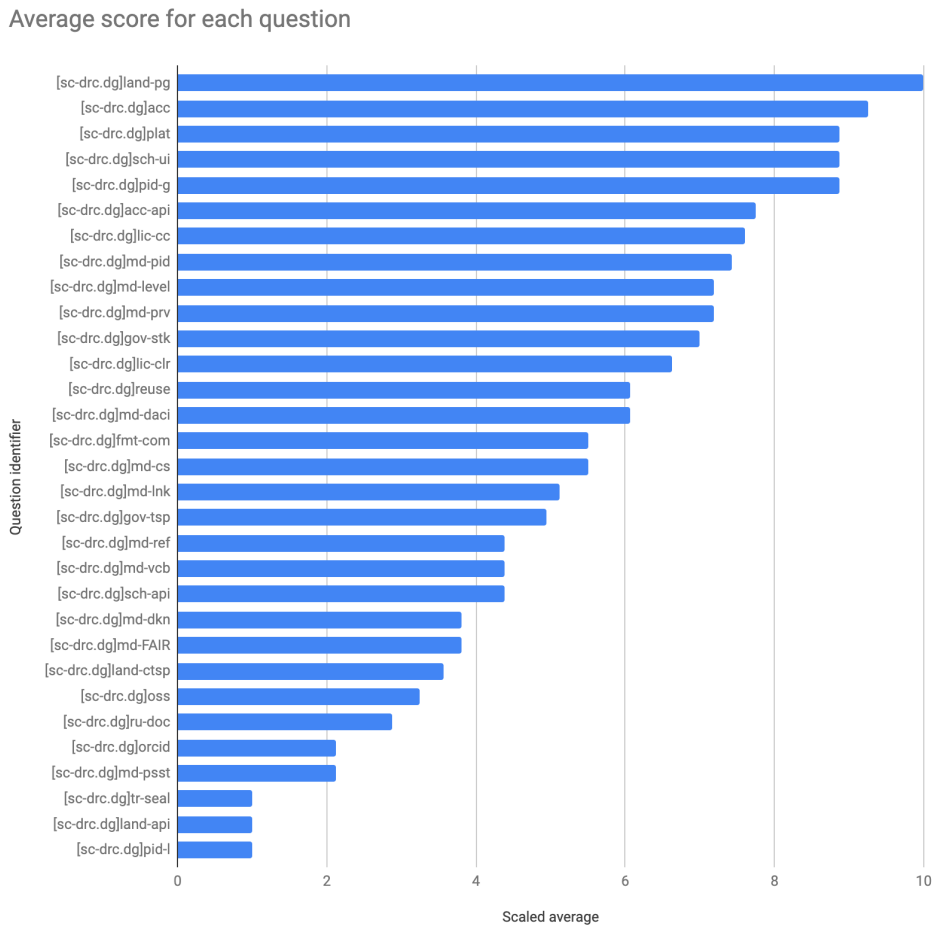

Figure 3: Average scaled score for each question across all repositories. Questions are ordered on the Y axis according to highest average score (top) to lowest score (bottom). The data underlying the figure is available in Bar-Sinai et al., 2020 in the summary-transcript.tsv file. The average scaled score was calculated per question and then the results were sorted from highest to lowest.

The answers to these questions are used to assign OFCT properties and flags in the Policy Space. Flags represent a binary rating; if the flag is assigned, then the repository meets that criterion, e.g., openFormat means that the repository makes data available in an open format. The properties and the flags assigned by the PolicyModels software and their meaning are provided in Table 4.

Our instrument calculates an overall rating per OFCT dimension, as shown in Figure 4. For a repository to be rated fully compliant, it would have to receive an acceptable score for all dimensions that evaluate that principle; conversely to be rated non-compliant would require an unacceptable score on all dimensions. This calculation is performed using PolicyModels, and is based on the range of acceptable and unacceptable values in various dimensions of the instrument's policy space. Note that we do not provide scores for individual repositories in this paper, as our intent is not to grade them. However, the completed questionnaires for the individual repositories are available in (Michael Bar-Sinai, Murphy, and Martone 2020).

As seen in Figure 4, at least one repository scored as fully compliant in each of the Open, Findability, Accessibility, Reusability and Citability dimensions. Conversely, three repositories received the lowest rating for Findability and one for Citability. No single repository was equally good - or bad - on all dimensions, that is, the same repositories did not receive either all of the highest or lowest scores. The most flags assigned to a single repository was 15 while the fewest was 5.

Table 4: Ratings for each OFCT property and flag

Properties and flags	Repository counts	QID	Short Explanation
Open			
Restrictions	none:6 minimal:2 significant:0	acc	Level of restrictions imposed by the repository in order to access datasets.
CCLicenseCompliance	nonCompliant:0 none:3 adequate:0 good:4 full:1	lic-cc	Commons-compliance level of the repository license
openFormat	no:4 yes:4	reuse	Is the data available in an open (non-proprietary) format?
platformSupportsData-Work	no:1 yes:7	plat	Does the repository platform make it easy to work with (e.g. download/re-use) the data?

ccLicenseOK	no:3 yes:5	lic-cc	Are the data covered by a commons-compliant license? (any answer except "closed" is considered a "yes")
restrictionsNotJustified	no:8 yes:0	acc	Does the repository impose "significant but not justified restrictions" on accessing the data?
FAIR:Findable			
PersistentIdentifier	none:1 internalPID:0 externalPID:7	pid-g, pid-l	Scope of persistent identifier assigned to the data, if any
IdInMetadata	none:1 partial:2 all:4	md-pid	Does the metadata clearly and explicitly include the identifier of the data it describes?}
MetadataGrade	minimal:0 limited:5 rich:3	md-level	Level of additional metadata that can be added to promote search and reuse of data
internalSearchOK	no:1 yes:7	sch-ui	Does the repository provide a search facility for the data and metadata?
FAIR:Accessible			
humanAccessible	no:1 yes:7	acc	Does the repository provide access to the data with minimal or no restrictions?
machineAccessible	no:2 yes:6	reuse, sch-api	Can the data be accessed by a computer? Note that this includes access both via UI and API, as web-based UI is by definition machine-accessible.
persistentMetadata	no:7 yes:1	md-psst	Does the repository have a policy that ensures the metadata (landing page) will persist even if the data are no longer available, either by policy or example?
licenseOK	no:3 yes:5	lic-clr	Does the repository provide a clear license for reuse of the data? (any answer except "no license")
stdApi	no:2 yes:6	acc-api	Can the (meta)data be accessed via a standards compliant API?
MetadataPersistence	no:7 byEvidence:0 byStatedPolicy:1	md-psst	Does the repository have a policy that ensures the metadata (landing page) will persist even if the data are no longer available?

FAIR:Interoperable			
MetadataFAIRness	minimal:4 allowed:3 enforced:1	md-FAIR	Do the metadata use vocabularies that follow FAIR principles?
StudyLinkage	none:0 freeText:6 textualMetadata:1 machineReadable-Metadata:1	md-lnk	Type of linkage between the published dataset and the paper that accompanied it
formalMetadataVocabularyOK	no:5 yes:3	md-vcb	Do the metadata use a formal, accessible, shared and broadly applicable language for knowledge representation?
fairMetadataOK	no:4 yes:4	md-FAIR	Do the metadata use vocabularies that follow FAIR principles? (any answer except "minimal")
qualifiedMetadataReferencesOK	no:4 yes:4	md-ref	Do the metadata include qualified references to other (meta)data? (any answer except "worst")
studyLinkageOK	no:6 yes:2	md-lnk	Linkage between the published dataset and the paper that accompanied it is "good" or "best".
MetadataReference-Quality	freeText:4 informal:2 formal:2	md-ref	Type of qualified references to other (meta)data, included in the (meta)data stored in the repository
FAIR:Reusable			
DocumentationLevel	lacking:4 adequate:3 good:1 full:0	ru-doc	Level of support offered by the repository for documentation that aids in proper (re)-use of the data
MetadataProvenance	unclear:0 adequate:5 full:3	md-prv	Are the (meta)data associated with detailed provenance?
documentationOK	no:4 yes:4	ru-doc	Does the repository require or support documentation that aids in proper (re)-use of the data? (any answer except "worst")
dkNetMetadataOK	no:5 yes:3	md-dkn	Does the repository accept metadata that is applicable to the dkNET community disciplines? (any answer except "worst")
communityStandard	no:4 yes:4	fmt-com	Does the repository enforce or allow the use of community standards for data format or metadata?
generalMetadata	no:4 yes:4	md-cs	Does the repository use a recognized community standard for representing basic metadata?

metadataProvenanceOK	no:0 yes:8	md-prv	Are the (meta)data associated with detailed provenance? (any answer except "worst")
DkNetMetadataLevel	none:5 dataset:1 datasetAndSubject:2	md-dkn	Does the repository accept metadata that is applicable to the dkNET community disciplines?
ReuseLicense	none:3 repositoryLevel:0 datasetLevel:5	lic-clr	Level at which the repository provides a clear license for reuse of the data
Citable			
MachineReadableLandingPage	none:1 exists:5 supportsDataCitation:2	land-pg, md-psst, land-api	Level of machine-readability of the dataset landing page (if any) provided by the repository
CitationMetadataLevel	none:2 partial:3 full:3	md-daci	Does the repository provide the required metadata for supporting data citation?
OrcidAssociation	none:6 supported:2 required:0	orcid	Does the repository allow the authors to associate their ORCID ID with a dataset?
Trustworthy			
GovernanceTransparency	opaque:2 partial:5 full:1	gov-tsp	Transparency level of the repository governance
SourceOpen	no:6 partially:0 yes:2	oss	Is the code that runs the data infrastructure covered under an open source license?
StakeholderGovernance	none:0 weak:2 good:2 full:2	gov-stk	Level of control stakeholders have in the repository's governance

Table 4: Overall ratings on each dimension measuring OFCT. Properties (Props) and Flags are assigned by the PolicyModels software based on the answers given. Properties are assigned at multiple levels depending on level of compliance, whereas all flags are binary and are only assigned if the repository meets the criteria. Repository Count = number of repositories with each rating; QID: ID of question that assigns the property/flag; Short explanation: meaning of the property or flag.

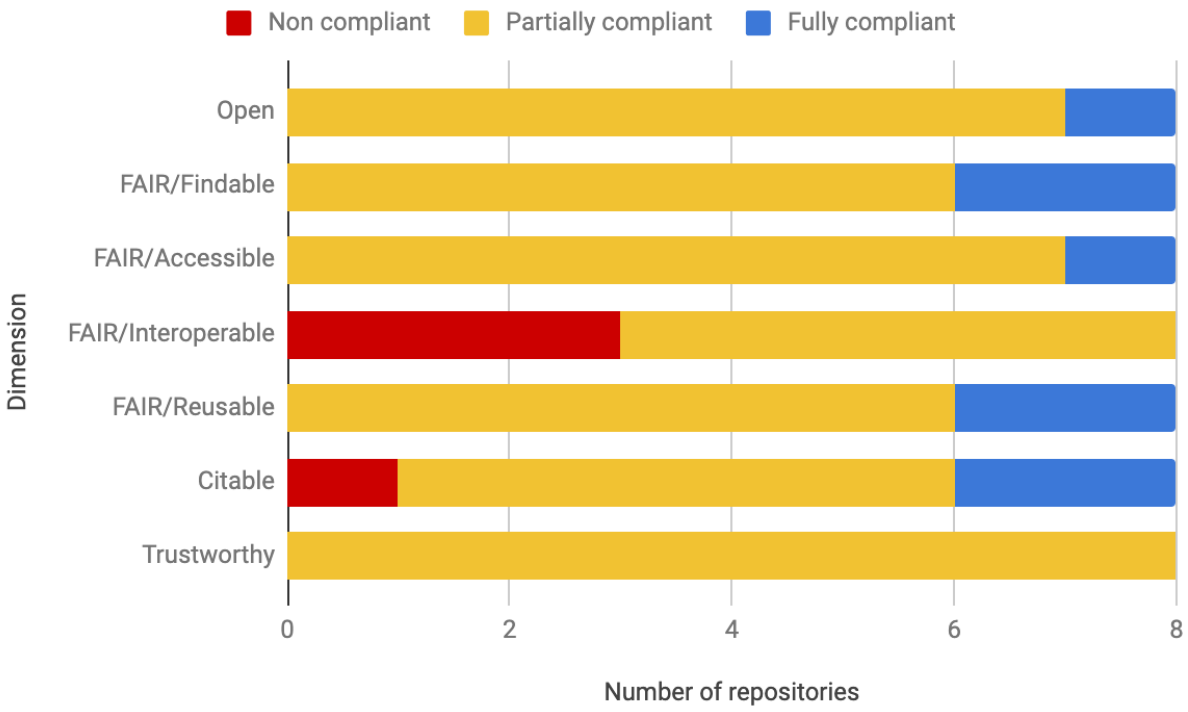

Figure 4: Overall ratings of repositories on OFCT criteria. The Y axis shows the individual dimensions and the X axis shows the number of repositories assigned each rating out of the 8 assessed. Red = Not compliant; Gold = Partially compliant; Blue = Fully compliant.

Open dimension

Seven repositories were scored as "Partially Open" and one as fully open (Figure 4) with details of the policy space for open criteria shown in Table 4. As biomedical repositories can deal with sensitive information that cannot be openly shared, they should adhere to the "As open as possible; as closed as necessary" principle. However, none of the repositories we evaluated had sensitive data and all were judged to make their data available with minimal to no restrictions, i.e., no approval process for accessing the data. We also evaluated repositories' policies against the open definition: "Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness." Thus, data have to be available to anyone, including commercial entities, and users must be free to share them with others. We thus examined the licenses against those

rated by the Open Knowledge Foundation as adhering to their definition (https://opendefinition.org/licenses/). One repository was considered fully compliant, 4 were rated as "good" with respect to open licenses, 3 had no licenses (Table 4; CCLicenseCompliance). The four rated as "good" did not receive the best score due to practices such as allowing the user to select from a range of licenses, some of which restricted commercial use.

FAIR dimension

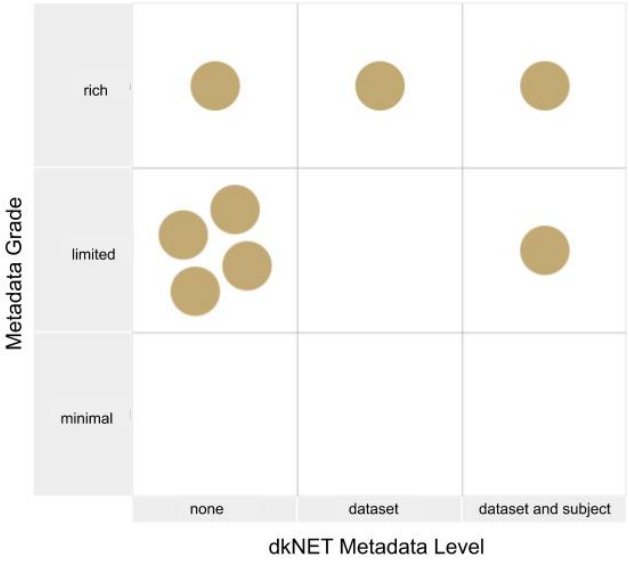

FIgure 5: Assessment of the degree of descriptive metadata (X) vs relevant biomedical metadata (dkNET Metadata Level) (Y). The Metadata Grade assesses whether the repository complies with the Findable principle for Rich Metadata, while the dkNET metadata measures the degree to which the repository supports the Reusable principle requiring "a plurality of relevant attributes". Relevance here was assessed with respect to dkNET. Only one repository received the highest score for both categories.

Our questions on FAIR evaluated both compliance with specific FAIR criteria, e.g., the presence of a persistent identifier or with practices that support FAIR, e.g., providing landing pages and providing adequate documentation to promote reuse. Evaluating a repository

against some principles also required that we define concepts such as “rich metadata” (FAIR principle F2) and a “plurality of relevant attributes” (FAIR principles R1).

Rich metadata were considered to comprise basic descriptive metadata, i.e., dataset title, description, authors but also metadata specific to biomedical data, e.g., organism, disease conditions studied and techniques employed (Q:**md-level**). “A plurality of relevant attributes” was defined in question **md-dkn** as providing sufficient metadata to understand the necessary context required to interpret a dkNET relevant biomedical dataset. Such metadata includes subject level attributes, e.g., ages, sex and weight along with detailed experimental protocols. Figure 5 positions each repository in the metadata policy space and shows that only one repository fully satisfied both metadata requirements.

Figure 4 shows that the majority of repositories were either partially or fully compliant with all the Findability and Accessibility dimensions. Two repositories achieved the highest rating in Findability. Seven out of the 8 repositories supported external PIDs, either DOIs or accession numbers registered to identifiers.org. One repository issued no identifiers. Only 1 repository was considered fully accessible because only 1 repository had a clear persistence policy (Q:md-psst). Both the JDDCP and FAIR principles state that metadata should persist even if the accompanying data are removed. We considered either an explicit policy or clear evidence of such a practice as acceptable, e.g., a dataset that had been withdrawn but whose metadata remained.

Overall scores were lowest for the interoperability dimensions, with 3 repositories being judged non-interoperable. Only one of the repositories achieved the StudyLinkage flag which indicated that they had fully qualified references to other data, in other words, that the relationship between a metadata attribute and a value was both machine readable and informative. We measured this property by looking at how repositories handled supporting publica-

tions in their metadata, e.g., did they specify the exact relationship between the publication and the dataset? To measure this, we looked at the web page markup ("view source") and also checked records in DataCite.

Two repositories achieved the highest score for reusability, while the remainder were considered partially reusable. Five repositories were judged as having inadequate metadata for providing experimental context, 4 as having inadequate user documentation, while 3 did not provide a clear license.

Fig 6: Repositories plotted against two dimensions of data citation. The Y axis shows support for citation metadata and the X axis for ORCID support. Two repositories support ORCID and provide full citation metadata. Two repositories have no support for data citation and the others have partial support.

Citable dimension

Data citation criteria included the availability of full citation metadata and machine-readable citation metadata according to the JDDCP ((Starr et al. 2015);(Fenner et al. 2019); (Cousijn et al. 2018)). We also evaluated the use of ORCIDs, as linking ORCIDs to datasets facilitates assigning credit to authors. As shown in Figure 5, only two repositories supported ORCID and provided full citation metadata. Consequently, 2 repositories were judged to fully support data citation, while the remainder were judged as partially (N=5) or not supporting (N=1) data citation. Many of the repositories had a citation policy, but most of these policies requested citation of a paper describing the repository and contributor of the data acknowledged rather than creating full citations of a particular dataset. Two were judged not to have sufficient metadata to support full citation, e.g., listing only the submitter and not other authors [see question med-daci].

Trustworthy dimension

Trustworthiness was largely assessed against the Principles of Open Infrastructures (Bilder et al., 2015) and the CoreTrustSeal criteria. The questionnaire originally probed the different certification criteria recommended by the CoreTrustSeal but we dropped this approach in favor of a single binary question on whether or not the repository was certified by CoreTrustSeal or equivalent. If a repository was certified, it would automatically be rated fully trustworthy. However, none of the eight repositories provided evidence of such a certification.

In accordance with the Principles of Open Infrastructures, we measured the degree to which the governance of the repository was transparent and documented and whether the repository was stakeholder governed. Only one repository received the highest rating for each of these, while 1 had virtually no information on how the repository is governed, e.g., who is the owner of the repository, or how decisions are made. Although 6 of the repositories were re-

searcher-led, it wasn't always clear how the stakeholder community was involved in oversight, e.g., a scientific advisory board. Finally, the Principles of Open Infrastructures recommends that the software underlying the repository be open source, so that if the repository ceases to be responsive to the community, it could be forked. Two of the repositories provided links to a GitHub repository with a clear open source license.

Discussion

As part of dkNET.org's efforts to promote data sharing and open science, we undertook an evaluation of current repositories supporting research domains of relevance to dkNET. Our ultimate goal is to provide tools to help researchers within these domains select an appropriate repository for their research data. Some of the data acquired with this instrument will be used to enhance dkNET's repository listings with information that might be important to a researcher when selecting a repository, e.g., does the repository support data citation. We also want to serve as a resource for those developing new dk data repositories by defining a set of important functions such repositories should support. More attention is now being paid in biomedicine to certification instruments such as the CoreTrustSeal, as evidenced by a newly released RFA for data repositories by the US National Institutes of Health (NIH 2020).

A good-faith effort was made to try to answer the questions accurately, although reviewing biomedical repositories is challenging. Each of the sites is organized differently and the specialized research repositories were developed to serve different communities and use cases. Therefore, to evaluate specific dimensions required significant engagement with the site, even in some cases requiring us to establish accounts to see what metadata was gathered at time of upload. Discovery of these types of routes, e.g., that ORCIDs are only referenced

when you establish an account, required us often to go back and re-evaluate the other repositories using this same method.

Only two of the repositories gave any indication that their functions or design were informed by any of the OFCT principles, specifically mentioning FAIR. The lack of explicit engagement with these principles is not surprising given that most of the repositories were established before these principles came into existence. For this reason, we gave credit for what we called "OFCT potential" rather than strict adherence to a given practice. We used a sliding scale for many questions that would assign partial credit. For example, if the repository did have landing pages at stable URLs we gave them some credit, even if the identifier was not strictly a PID. Such IDs could easily be turned into PIDs by registering them with a resolving service such as Identifiers.org or N2T.org (Wimalaratne et al. 2018).

In addition to finding relevant information, consistent scoring of the repository was also a challenge. Principles are designed to be aspirational and to provide enough flexibility that they will be applicable across multiple domains. There is therefore a certain amount of subjectivity in their evaluation particularly in the absence of validated, established standards. For example, one of the repositories issued persistent identifiers at the project level but not to the data coming from the individual studies. In another website not included in the final evaluation sample, DOIs were available upon request. Are these considered compliant? One could argue both ways.

As described in the methods, we did not attempt to cover all aspects of the underlying principles, we selected those for which we could develop reasonable evaluation criteria. One very important issue covered by CoreTrustSeal, the newly published TRUST principles (Lin et al. 2020) and Principles of Open Infrastructure (Bilder, Lin, and Neylon 2015) is long term sustainability. Although critical, we do not think that an external party such as ourselves is in a posi-

tion to comment on the long term sustainability plan for a given repository. Long term sustainability for biomedical infrastructure is a known problem and one for which there are currently few concrete answers as support of most researcher-led infrastructures is in the form of time-limited grants. Our instrument is relevant to this issue, however, as OFCT practices such as FAIR, open formats, open software and good governance practices make repositories more likely to be sustainable as they facilitate transfer of data across organizations.

To our knowledge, this is the first evaluation instrument that was designed specifically around OFCT. However, since the issuance of the FAIR data principles, several initiatives have invested in the development of tools that are designed to assess the level of data FAIRness, including those that are meant to evaluate on-line data repositories. Some funders such as the EU and NIH are developing policies around FAIR data which may include a more formal assessment of FAIRness. Such tools include FAIRmetrics[1], FAIR Maturity Indicators (Wilkinson et al. 2019)), FAIRshake (Clarke et al. 2019) and the FORCE11/Research Data Alliance evaluation criteria (McQuilton et al. 2020).

The FAIR Maturity Indicators and FAIRshake toolkits differ from ours in that they are intended to employ either fully automated or semi-automated approaches for determining FAIRness. As we show here, some aspects of FAIR require interpretation, e.g., “a plurality of relevant attributes”, making it difficult to employ fully automated approaches. In the case of “rich metadata” and “plurality of relevant attributes”, dkNET is evaluating these based on our criteria, that is, the type of metadata we think are critical for biomedical studies in our domain. These may not be universal. On the other hand, automated tools for determining the level of machine readability for features such as landing pages would make evaluation much simpler than our current process. We will likely incorporate some of these tools into future versions of the instrument.

[1] https://github.com/FAIRMetrics/Metrics

While evaluation tools can be powerful, there are downsides to rushing into too rigid an interpretation of OFCT. First, communities are still coming together to determine what constitutes OFCT for their constituents and what can be reasonably implemented at this time. As noted in the introduction, data repositories have to straddle two worlds: providing traditional publishing/library functions to ensure findability and stability, while at the same fulfilling more traditional roles of scientific infrastructures for harmonizing and reusing data. Thus, evaluating a repository from a journal's perspective may not be the same as from a researcher's perspective.

Second, (Sansone et al. 2020) analyzed different evaluation metrics for data repositories and found that although they agree on some dimensions, they don't agree on all. Based on their analysis, they have made specific recommendations as to the types of functions they should support and the information which should be available. Such results indicate that it is still perhaps early days for understanding what constitutes best practices for a data repository across all disciplines. Our understanding of such practices may evolve over time as data sharing becomes more mainstream. As already noted, for example, early efforts in data sharing necessarily focused on deposition of data. Less attention, perhaps, was paid to what it takes for the effective reuse of the data. While the FAIR principles emphasize machine-readable attributes for achieving reusability without human intervention, some studies suggest that the human factor may be more critical for some types of data (Faniel and Yakel, 2017). For these types, having a contact person and an accompanying publication makes it much easier to understand key contextual details (Faniel and Yakel 2017; Turner et al. 2011). As we start to see more reuse of data, it may be possible to employ more analytical methods for determining best practices based on actual use cases.

For these reasons, we deliberately refrained from assigning grades or calling out individual repositories in the work presented here. (Wilkinson et al. 2019) noted that many repositories

which were evaluated early on using FAIRmetrics expressed resentment. We recognize the struggles that those who develop and host scientific data repositories undergo to keep the resource up and running, particularly in the face of uncertain funding. Generally, these repositories were founded to serve a particular community, and the community itself may not be demanding or engaging with OFCT principles. We therefore favor flexible approaches that allow individual communities to interpret OFCT within the norms of their community and not entirely according to the dictates of external evaluators. Nevertheless, research data repositories, after operating largely on their own to determine the best way to serve research data, are going to have to adapt to meet the challenges and opportunities of making research data a primary product of scientific research.

References

Bar-Sinai, Michael, Fiona Murphy, and Maryann Martone. 2020. *Codeworth-gh/dkNET-Decision-Trees: Paper Submission*. https://doi.org/10.5281/zenodo.4069491.

Bar-Sinai, M., L. Sweeney, and M. Crosas. 2016. "DataTags, Data Handling Policy Spaces and the Tags Language." In *2016 IEEE Security and Privacy Workshops (SPW)*, 1–8.

Berman, Helen M., Gerard J. Kleywegt, Haruki Nakamura, and John L. Markley. 2012. "The Protein Data Bank at 40: Reflecting on the Past to Prepare for the Future." *Structure* 20 (3): 391–96.

Bilder, Geoffrey, Jennifer Lin, and Cameron Neylon. 2015. "Principles for Open Scholarly Infrastructures." *Science in the Open*. https://cameronneylon.net/blog/principles-for-open-scholarly-infrastructures/.

Bosman, J.; Bruno, I; Chapman, C; Greshake Tzovaras, B; Jacobs, N; Kramer, B; Martone, M; Murphy, F; O'Donnell, D P; Bar-Sinai, M; Hagstrom, S., Utley, J., Veksler, L. (2017) The Scholarly Commons - principles and practices to guide research communication OSF Preprints, http://dx.doi.org/10.31219/osf.io/6c2xt

Clarke, Daniel J. B., Lily Wang, Alex Jones, Megan L. Wojciechowicz, Denis Torre, Kathleen M. Jagodnik, Sherry L. Jenkins, et al. 2019. "FAIRshake: Toolkit to Evaluate the FAIRness of Research Digital Resources." *Cell Systems* 9 (5): 417–21.

CoreTrustSeal Standards and Certification Board. 2019. *CoreTrustSeal Trustworthy Data Repositories Requirements: Glossary 2020–2022*. https://doi.org/10.5281/zenodo.3632563.

Cousijn, Helena, Amye Kenall, Emma Ganley, Melissa Harrison, David Kernohan, Thomas Lemberger, Fiona Murphy, et al. 2018. “A Data Citation Roadmap for Scientific Publishers.” *bioRxiv*. https://doi.org/10.1101/100784.

Data Citation Synthesis Group. 2013. “Joint Declaration of Data Citation Principles.” https://www.force11.org/datacitationprinciples.

Faniel, I. M., and E. Yakel. 2017. “Practices Do Not Make Perfect: Sharing, Disciplinary Data; Practices, Reuse.” In *Curating Research Data Volume One: Practical Strategies for Your Digital Repository*, edited by L. R. Johnson, 1:103–25. Association of College and Research Libraries.

Fenner, Martin, Mercè Crosas, Jeffrey S. Grethe, David Kennedy, Henning Hermjakob, Phillippe Rocca-Serra, Gustavo Durand, et al. 2019. “A Data Citation Roadmap for Scholarly Data Repositories.” *Scientific Data* 6 (1): 28.

Lin, Dawei, Jonathan Crabtree, Ingrid Dillo, Robert R. Downs, Rorie Edmunds, David Giaretta, Marisa De Giusti, et al. 2020. “The TRUST Principles for Digital Repositories.” *Scientific Data* 7 (1): 144.

Martone, Maryann E., Fiona Murphy, and Michael Bar-Sinai. 2020. *dkNET Repository Compliance*. https://doi.org/10.5281/zenodo.4086039.

McQuilton, Peter, Susanna-Assunta Sansone, Helena Cousijn, Matthew Cannon, Wei Chan, Ilaria Carnevale, Imogen Cranston, et al. 2020. “FAIRsharing Collaboration with DataCite and Publishers: Data Repository Selection, Criteria That Matter.” Open Science Framework. https://doi.org/10.17605/OSF.IO/N9QJ7.

NIH. 2020. “PAR-20-089: Biomedical Data Repository.” Accessed October 14, 2020. https://grants.nih.gov/grants/guide/pa-files/PAR-20-089.html.

OpenAire. 2020. “How to Select a Data Repository?” Accessed October 13, 2020. https://www.openaire.eu/opendatapilot-repository-guide.

Open Knowledge Open Definition Group. 2020 “The Open Definition - Open Definition - Defining Open in Open Data, Open Content and Open Knowledge.” Accessed February 25, 2020.

https://opendefinition.org/.

Sansone, Susanna-Assunta, Peter McQuilton, Helena Cousijn, Matthew Cannon, Wei Mun Chan, Sarah Callaghan, Ilaria Carnevale, et al. 2020. “Data Repository Selection: Criteria That Matter,” October. https://doi.org/10.5281/zenodo.4084763.

Starr, Joan, Eleni Castro, Mercè Crosas, Michel Dumontier, Robert R. Downs, Ruth Duerr, Laurel L. Haak, et al. 2015. “Achieving Human and Machine Accessibility of Cited Data in Scholarly Publications.” *PeerJ Computer Science* 1 (May): e1.

Turner, Charles F., Huaqin Pan, Gregg W. Silk, Mary-Anne Ardini, Vesselina Bakalov, Stephanie Bryant, Susanna Cantor, et al. 2011. “The NIDDK Central Repository at 8 Years--Ambition, Revision, Use and Impact.” *Database: The Journal of Biological Databases and Curation* 2011 (January): bar043.

Whetzel, Patricia L., Jeffrey S. Grethe, Davis E. Banks, and Maryann E. Martone. 2015. “The NIDDK Information Network: A Community Portal for Finding Data, Materials, and Tools for Researchers Studying Diabetes, Digestive, and Kidney Diseases.” *PloS One* 10 (9): e0136206.

Wilkinson, Mark D., Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” *Scientific Data* 3 (March): 160018.

Wilkinson, Mark D., Michel Dumontier, Susanna-Assunta Sansone, Luiz Olavo Bonino da Silva Santos, Mario Prieto, Dominique Batista, Peter McQuilton, et al. 2019. “Evaluating FAIR Maturity through a Scalable, Automated, Community-Governed Framework.” *Scientific Data* 6 (1): 174.

Wimalaratne, Sarala M., Nick Juty, John Kunze, Greg Janée, Julie A. McMurry, Niall Beard, Rafael Jimenez, et al. 2018. “Uniform Resolution of Compact Identifiers for Biomedical Data.” *Scientific Data* 5 (May): 180029.

Acknowledgements

This work was supported by NIH grant# 3U24DK097771-08S1 from the National Institutes of Diabetes and Digestive and Kidney Diseases. The original decision tree was developed through the FORCE11 Scholarly Commons working group supported by an award from The Leona M. and Harry B. Helmsley Charitable Trust Biomedical Research Infrastructure Program to FORCE11. The authors wish to thank Drs. Ko Wei Lin and Jeffrey Grethe for helpful comments.

Author Contributions

All three authors contributed substantially to the design, execution and writing of the study. MBS performed all of the coding of the instrument in PolicyModels and the visualization tool. He created the summary statistics. FM and MM performed the evaluation of the repositories.

Competing Interests

Dr. Martone is on the board and has equity interest in SciCrunch Inc., a tech startup that develops tools and services in support of Research Resource Identifiers.
Dr Murphy is on the board of Dryad Data Repository.

Supplemental material

Supplemental table S1
Question ids and their abbreviated meaning. PolicyModels allows specifying an id for each question. This id is later used to identify that question, and to localize its text. The ids we use here pertain to the subject of the question. The table below explains each abbreviation.

`acc`	Provide access to the data
`acc-api`	Provide access to the data via API
`fmt-com`	Data format - allow community standards
`gov-stk`	Governance - stakeholders involvement

gov-tsp	Governance - transparency
land-api	Landing page - machine readability
land-ctsp	Landing page - data citation support
land-pg	Landing page - pointed by PID
lic-cc	License - Creative Commons compliance
lic-clr	License - clarity
md-FAIR	Metadata - FAIR compliance
md-cs	Metadata - using community standards
md-daci	Metadata - supporting data citation
md-dkn	Metadata - applicability to dkNET community
md-level	Metadata - richness level
md-lnk	Metadata - linking to publication
md-pid	Metadata - includes PID
md-prv	Metadata - includes provenance
md-psst	Metadata - persistence (even after the data is gone)
md-ref	Metadata - qualified references to other (meta)data
md-vcb	Metadata - vocabulary usage
orcid	ORCID association support
oss	Open-Source infrastructure
pid-g	PID - using a global PID (e.g. DOI)
pid-l	PID - local (e.g. self-assigned)
plat	Platform for working with the data
reuse	Reuse - licensing
ru-doc	Reuse - supporting documentation
sch-api	Search and access via API
sch-ui	Search and access via UI
tr-seal	Core Trust Seal support