



# State of the Data: Assessing the FAIRness of US Geological Survey Data

RESEARCH PAPER

VIVIAN B. HUTCHISON

TAMAR NORKIN

LISA S. ZOLLY

LESLIE HSU

\*Author affiliations can be found in the back matter of this article

ubiquity press

## ABSTRACT

In response to recent shifts towards open science that emphasize transparency, reproducibility, and access to research data, the US Geological Survey (USGS) conducted a study to assess the degree to which USGS data assets meet the FAIR data principles (Findable, Accessible, Interoperable, and Reusable). The USGS designed and applied a methodology for quantitative analysis of FAIR characteristics. A new rubric was derived from a crosswalk of existing FAIR evaluation frameworks and customized for the USGS. The rubric, consisting of 62 yes/no questions, was applied to 392 metadata records of USGS data products published between 1987 and 2022. Results were analyzed to show which FAIR characteristics were most and least present in the metadata and how these scores changed after the implementation of data policy requirements in 2016. Aggregated scores showed specific areas of strength and needed improvements. The greatest increases in FAIR scores over time were for elements that were required by new data policies, especially in the 'Findable' category. Based on the results, this paper presents strategies to further improve USGS alignment with FAIR. The suggested strategies are organized in four key areas: USGS data repository characteristics, training and communities of practice, data management policy considerations, and metadata standards, tools, and best practices.

## CORRESPONDING AUTHOR:

**Vivian B. Hutchison**

US Geological Survey, Science Analytics and Synthesis, Denver, Colorado, USA

[vhutchison@usgs.gov](mailto:vhutchison@usgs.gov)

## KEYWORDS:

FAIR Principles; FAIR data; open data; research data management

## TO CITE THIS ARTICLE:

Hutchison, V B, Norkin, T, Zolly, L S and Hsu, L 2024 State of the Data: Assessing the FAIRness of US Geological Survey Data. *Data Science Journal*, 23: 22, pp. 1–20. DOI: <https://doi.org/10.5334/dsj-2024-022>

Recent shifts towards open science in both government and research spheres emphasize transparency, reproducibility, and public access to research data (OSTP 2022; Ramachandran et al. 2021). In the United States, the 2018 *Foundations for Evidence-Based Policymaking Act*, or Evidence Act (US Congress 2018), codified into law recent open data directives, including the 2013 *Open Data Policy – Managing Information as an Asset* (OMB 2013) and the 2017 *OPEN Government Data Act* (US Congress 2017). Additionally, Federal science organizations with a research budget greater than \$100 million must develop a public access plan to guide implementation of open science directives (OSTP 2013). Collectively, these Federal policies have led to the widespread adoption of open data practices across government agencies.

The US Geological Survey (USGS) has a long history of making scientific data available to stakeholders and the public (FSPAC 2011). An early and consistent proponent of public access to data, the USGS publishes information products including scientific reports, data releases, real-time data streams, and online web tools. These products support scientific decision-making and facilitate reuse of data beyond their original purpose.

The USGS has established internal policies and workflows to help its scientists meet government requirements and has shifted its research culture to integrate data management activities into daily practices. The USGS *Fundamental Science Practices* (FSP) define comprehensive policies and procedures for ensuring the quality and integrity of USGS science (FSPAC 2011; FSPAC 2023). In 2016, the USGS added policies to the FSP that formally require the management of scientific data as strategic assets and mandate the public release of all scientific data supporting scholarly conclusions in publications (USGS 2017a; USGS 2017b; USGS 2017c; USGS 2017d). The 2016 USGS Public Access Plan, *Public Access to Results of Federally Funded Research at the US Geological Survey*, outlines the steps that the USGS planned to take to operationalize open data policies and best practices (USGS 2016).

Introduced by an international consortium of researchers and organizations in 2016, the FAIR Principles (Wilkinson et al. 2016) provide a framework for making scientific data Findable, Accessible, Interoperable, and Reusable. The FAIR principles can provide a concise and measurable strategy for improving data management practices and enabling effective data discovery and reuse. Organizations around the world are promoting and evaluating alignment with the FAIR Principles (for example, Clarke et al. 2019; FAIRsFAIR 2022; Jones et al. 2019; Peng et al. 2023; RDA 2020; Wilkinson et al. 2018). The current landscape and the need for harmonizing the different methods is well-described in Peng (2023) and references therein.

Two recent USGS projects focused on applications of the FAIR principles to data practices. The first was a workshop in 2019 that brought together data professionals from across the USGS to discuss actions that could improve alignment with FAIR. The resulting report, *Opportunities to Improve Alignment with the FAIR Principles for US Geological Survey Data* (Lightsom et al. 2022), contains more than 100 proposed strategies.

The second, the State of the Data project, is the focus of this paper. While the FAIR Workshop generated strategies based on workshop discussions, the State of the Data project sought to quantify and evaluate alignment with FAIR using standardized dataset assessments. The State of the Data project team created a peer-reviewed rubric that can be used to evaluate individual datasets against itemized elements of FAIR (Hutchison et al. 2023). The team applied the rubric to a sample set of 392 USGS data products and calculated FAIR scores based on the degree to which the datasets align with the FAIR principles. The sample set selection method is described in the ‘Methods’ section of this paper.

Two overarching goals guided the State of the Data project:

1. Develop a methodology for a quantitative analysis of the FAIR characteristics of USGS data and determine a baseline status for the current overall FAIRness of USGS data.
2. Based on the results of the analysis, propose strategies for how the USGS can improve its alignment with FAIR. The original methodology can be reused in the future to measure progress.

In addition, the following research questions informed the project's methodology:

- To what degree have the recent USGS data policies affected compliance with the FAIR principles? That is, are there measurable differences between data published before and data published after the institution of Fundamental Science Practices for data management in FY2016?
- How well did the project's results support observational understandings of the strengths and weaknesses of USGS practices? Do the project's suggested strategies align with those described in *Opportunities to Improve Alignment with the FAIR Principles for US Geological Survey Data* (Lightsom et al. 2022), which were generated following discussions among USGS data management experts?

## METHODS

### OVERVIEW

The State of the Data project was conducted in two phases. The first focused on planning, methodology development, and a pilot project. Lessons from the pilot project were used to refine the methodology for the second phase, in which a custom rubric was developed and used to assess a sample set of public USGS data products.

Additional methodology details are available in the metadata of the associated data release (Hutchison et al. 2023).

The first phase of the project was conducted in 2020. We studied existing matrices and evaluation techniques for concepts including data maturity, AI-readiness, and FAIRness. Two matrices were selected to be tested in a phase one pilot project: National Oceanic and Atmospheric Administration's (NOAA) Data Stewardship Maturity Matrix (DSMM) (Peng n.d.; Peng et al. 2015) and a draft AI-Readiness Matrix (Office of Science Technology and Policy (OSTP) Subcommittee on Open Science, unpublished data, 2019).<sup>1</sup> A sample set of 163 public datasets was selected at random from ScienceBase, a USGS data repository (Hutchison et al. 2021), and evaluated against the DSMM and AI-Readiness Matrix.

In addition, a selection of FAIR evaluation frameworks was aggregated and organized into a crosswalk, to more comprehensively sample work already completed by various data communities (Clarke et al. 2019; Go FAIR n.d.; Habermann and Jones 2020; Jones et al. 2019; RDA 2020). The crosswalk identified and categorized elements that were shared across the evaluation frameworks.

In 2021, the sample set for the phase two evaluation was selected from across the USGS, rather than a single repository. We collected the sample set from the USGS Science Data Catalog (SDC), a metadata catalog that aggregates metadata from USGS science centers and programs (USGS n.d.-b). The purpose was to generate data that can be used to analyze USGS-wide trends.

The scope of this evaluation focused on the FAIR principles, instead of on the broader concepts of analysis-readiness, AI-readiness, and general data maturity. This decision was based on the need to evaluate a wide variety of datasets in a consistent and quantitative way. The scope was also influenced by existing USGS data guidance and resources. These include data and metadata checklists that support the required internal review process (USGS n.d.-a). The new rubric is intended to complement, not duplicate, these resources, so it addresses FAIR characteristics at a coarser level than those addressed in a comprehensive peer review.

### RUBRIC CREATION

The USGS rubric was built from the crosswalk of FAIR framework evaluations from phase one. We took the list of characteristics from the crosswalk and converted it into a list of questions that have possible answers of 'Yes,' 'No,' and 'Not Applicable.' The list of questions was then customized for USGS practices and policies. The vast majority of USGS metadata records are in the Content Standard for Digital Geospatial Metadata (CSDGM) format (FGDC 1998); therefore, the rubric's scoring guidance was specifically tailored for this format.

---

<sup>1</sup> At the time of publication, the AI-Readiness Matrix was not available from the OSTP Subcommittee on Open Science.

We categorized the questions based on their level of importance for the FAIR principles, using the terms *essential*, *intermediate*, and *advanced*. Essential questions are necessary for FAIR data in the USGS. Intermediate questions are important and beneficial but not always necessary for every USGS data release. Advanced questions are useful and add FAIR value but may not be applicable or available for all USGS data releases.

Many decisions that impact the production of FAIR data are controlled by data authors; for example, decisions to use machine-readable file formats. Other FAIR characteristics, however, are determined by the repositories that host the data; for example, enabling machine or application programming interface (API) access to the data. We decided to separate the characteristics that were dependent on repository capabilities from the other characteristics in the rubric, so that data authors and managers could focus on the elements that are within their control. The set of repository-dependent characteristics are now listed in a separate tab in the rubric. Although they are important considerations for USGS data repository managers and decision makers, they may not be relevant considerations for data authors and reviewers.

The FAIR rubric is in Microsoft Excel Workbook format ([Hutchison et al. 2023](#)). There are separate tabs for each of the four categories of FAIR: Findable, Accessible, Interoperable, and Reusable. Scores are entered directly in the Workbook, which contains formulas that automatically calculate overall FAIR scores and populate a scorecard. After completion of the project, scores from all completed Workbooks were aggregated into a single tabular dataset ([Hutchison et al. 2023](#)).

Although Excel is not an open format, we decided to use it for data collection due to its availability, ease of use, and human readability. We also anticipated being able to use a script at the end of the process to parse the Excel files, aggregate the data, and share our results in an open, machine-readable format.

## RUBRIC REFINEMENT

The rubric used for the phase two assessments went through multiple rounds of updates, based on reviews, calibration exercises, and quality control checks of data assessments.

The first draft of the rubric was peer reviewed by USGS data managers. We later organized a workshop with a small group of USGS data managers, many of whom review data and metadata for their science centers. Workshop attendees applied the draft rubric to a data release and discussed the process. We used their input and suggestions to inform updates and improvements to the rubric. We also expanded the rubric's ReadMe tab to answer key questions and clarify points of confusion.

The rubric was then rigorously tested through several rounds of calibration checks to see if different assessors could score datasets in a consistent way. A group of eight USGS assessors scored the same five datasets independently and then compared scores. Questions with divergent answers were highlighted, measured, and discussed within the group to resolve differences. These calibration tests revealed significant discrepancies and led to the addition of two columns of scoring guidance within the rubric. The first described the specific conditions that define a score of 1, 0, and N/A. The second contained the specific fields in the CSDGM metadata standard that assessors should check to find their answers.

Throughout the dataset assessment process, we conducted regular spot checks on scores for consistency. Two members of the team independently assessed datasets that had already been assessed by others, and then compared the scores. Discrepancies were discussed and then reconciled. In some cases, clarifications or additional specifications were added to the scoring guides.

Because assessors enter their scores directly in the rubric's Excel Workbook, the order of columns can affect ease of use. A usability study was conducted to determine the order of columns that would optimize user experience.

After the assessment period was completed, we followed up again with the attendees of the data manager workshop. Attendees used the final version of the rubric, which included the revised scoring guidance, to evaluate a dataset. We then compared their results to evaluate consistency and updated the scoring guidance as needed.

## SELECTION OF DATASETS

We selected 392 metadata records from the USGS Science Data Catalog (SDC). Metadata records in the SDC describe data hosted across numerous data repositories, both internal and external

to the USGS. The USGS has approximately 90 science centers or programs with metadata records cataloged in the SDC. We selected metadata records from each of these centers, with the goal of creating a sample set that represents a cross section of USGS data products.

Up to six records were selected from each center's metadata collection. Metadata records were categorized by year for each center, and we randomly selected up to three that were published before 2017 and up to three that were published in 2017 or later. This allowed us to examine changes in FAIR scores in the periods before and after USGS data policies went into effect (USGS 2017a; USGS 2017b; USGS 2017c; USGS 2017d).

Each selected metadata record represents one USGS data release. The term 'data release' refers to the USGS publication format for digital data files, metadata files, and other supporting documentation, as described in the USGS Survey Manual chapter 502.8 (USGS 2017b). Some data releases contain multiple metadata records. In these cases, we evaluated only the part described by the selected metadata record.

We originally selected 400 metadata records from the SDC. Some records in the original sample set were removed due either to content type (e.g., metadata records associated with software releases instead of data releases) or redundancy (metadata records selected from the same data release). Most of these were replaced with other records from the same center and publication data ranges. In some cases, however, no other records were available from the same selection category. As a result, the total number of datasets assessed was slightly less than the original 400.

## SAMPLE SET COMPOSITION

The sample set composition was designed to provide a representative view across USGS public data products; however, due to the large and complex structure of the USGS, a comprehensive representation was not achievable for this project. The following are contributing factors.

The ScienceBase data repository is the largest USGS repository by total holdings and has the most metadata records in the SDC, so it is the repository most represented in the sample set. The USGS Water Mission Area's National Spatial Data Infrastructure (NSDI node) has the second most metadata records in the SDC and the sample set. At the time of the study, the NSDI node had its own website for data and metadata distribution. NSDI data products have since been moved into ScienceBase, so some FAIR scores may have changed (for example, the scores relating to landing page content).

The number of data releases published by individual science centers varies widely, often based on differences in center size and productivity. Some centers have fewer than six metadata records cataloged in the SDC, so fewer than six were selected for the sample set. As a result, not all centers are equally represented in the study. In addition, the number of published datasets in the USGS has increased over time, especially after recent USGS data policies (USGS 2016), so datasets published in 2017 or later are overrepresented in the sample set.

Lastly, all the metadata pulled from the SDC is in the CSDGM format. The USGS has published a relatively small number of data releases with ISO metadata records, and these were not included in the sample set.

## ASSESSMENT PROCESS

The first assessments completed were those used for calibration exercises. These were included in the final dataset. Subteams were created for the rest of the assessments. Each subteam consisted of a team lead and two or three assessors. The leads were responsible for conducting quality control checks of the assessments and answering questions about the scoring process. Following the calibration period, five assessors completed the assessments for the datasets in the sample set.

## DATA PROCESSING

Each assessment file contained the answers to 62 rubric questions. To prepare the assessments for analysis, we calculated the scores listed in Table 1. The 62 questions were organized in categories with different numbers of questions:

	NUMBER OF QUESTIONS
Total FAIR score	62
Findable score	24
Accessible score	8
Interoperable score	18
Reusable score	12
Essential score	37
Intermediate score	15
Advanced score	10

**Table 1** List of score categories, showing the number of questions in each. ‘FAIR’: Findable, Accessible, Interoperable, Reusable. For a list of all questions, see ‘Supplemental File 1: FAIR Rubric Questions’.

To compare scores for categories with different numbers of questions, we normalized all scores to a maximum of 100, taking into account the number of ‘Not Applicable’ answers. This means that each score can reach a maximum of 100, even if some of the rubric questions are not applicable.

$$\text{Score}_{\text{normalized}} = \text{score} / (\text{n\_questions} - \text{n\_NA}) * 100$$

The processing scripts and resulting data are available in the associated data release (Hutchison et al. 2023).

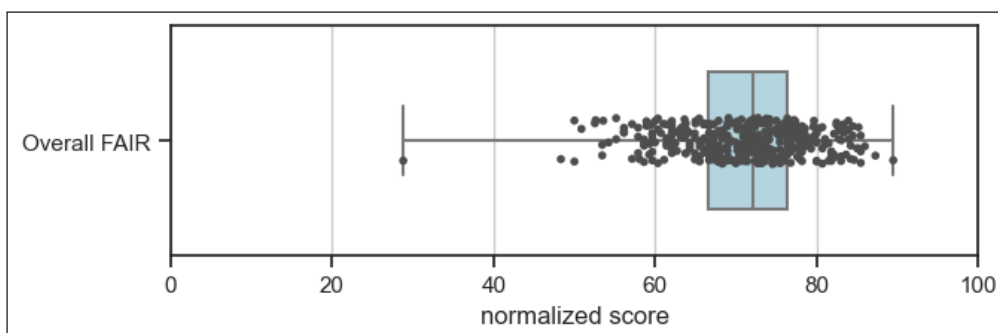
## RESULTS

The following results summarize the findings from the 392 dataset assessments. Interpretations of these results, in relation to USGS practices and policies, are described in the ‘Discussion’ section.

### OVERALL FAIR SCORES

The overall FAIR scores represent the number of relevant ‘Yes’ and ‘No’ answers for each of the 62 rubric questions. The results presented here use the normalized scores scaled to a maximum of 100 and do not penalize scores for questions that are not applicable, as described in the data processing section.

The distribution of overall scores is shown in Figure 1. The mean score is 71, with half of the assessments falling between 67 and 76. The minimum score was 29 and the maximum score was 90. Examples of the top scoring datasets are shown in Supplemental File 3: Example Datasets.

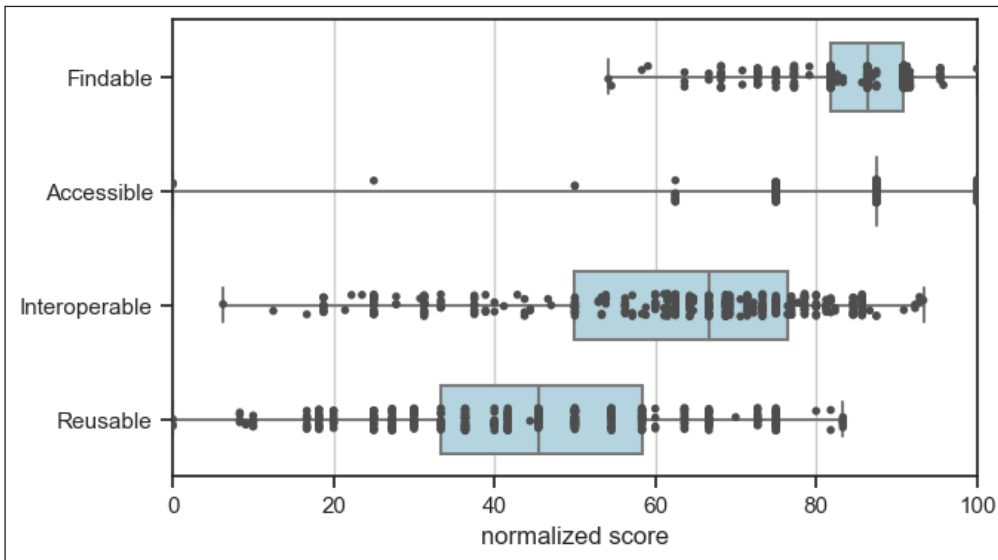


**Figure 1** Horizontal box plot with overlaid data points showing all 392 overall Findable, Accessible, Interoperable, and Reusable (FAIR) scores. Each score is normalized to a maximum of 100 and does not take into account questions that are not applicable.

### SCORES FOR FINDABLE, ACCESSIBLE, INTEROPERABLE, AND REUSABLE

Each overall FAIR score was broken down into four scores, one for each of the FAIR categories, also normalized to a maximum of 100 (Figure 2).

Findable and Accessible scores were the highest, with a mean of 86 and 85, respectively. Interoperable and Reusable mean scores were lower, with a mean of 62 and 45, respectively. Interoperable and Reusable scores also had a larger range, meaning that there was more variation between the datasets for how many questions in these categories received a score of ‘1’.



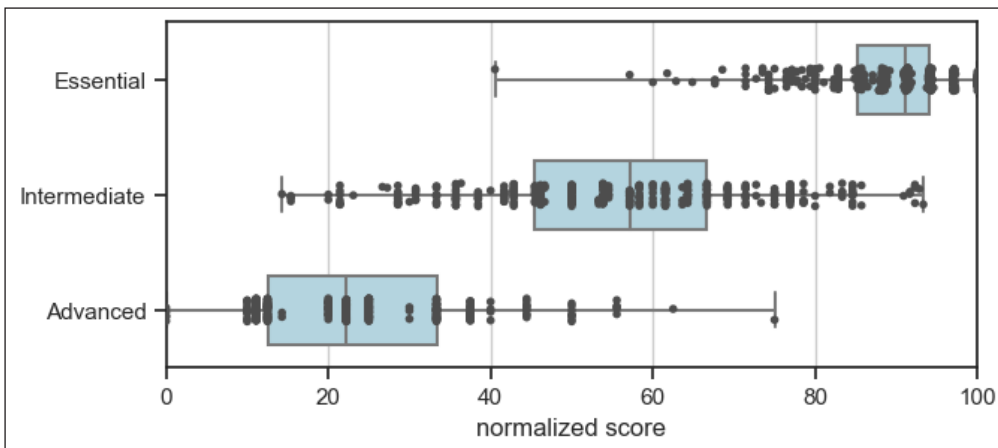
**Figure 2** Horizontal box plot with overlaid data points showing scores for all 392 assessments, broken down in the four FAIR principles: Findable, Accessible, Interoperable, and Reusable. Each score is normalized to a maximum of 100 and does not take into account questions that are not applicable.

### ESSENTIAL, INTERMEDIATE, ADVANCED SCORES

Each overall FAIR score can be broken down into the three designated levels of importance: Essential, Intermediate, and Advanced (Figure 3).

Scores for the Essential questions had a mean of 89, reflecting high percentage of ‘Yes’ answers. Intermediate category questions had a lower mean score of 56, and Advanced category questions had an even lower mean score of 24.

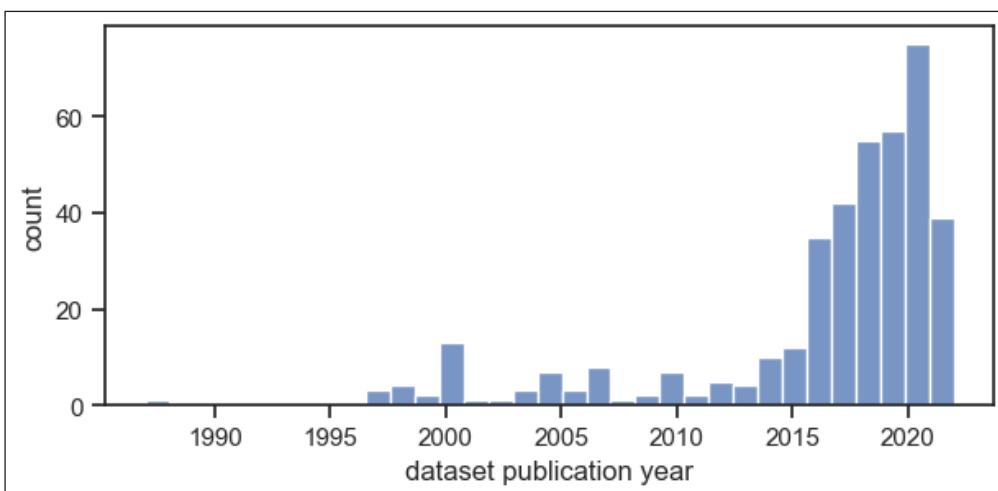
Intermediate and Advanced category questions may not be relevant to all datasets, but their lower scores indicate that there are areas for improvement.



**Figure 3** Horizontal box plot with overlaid data points showing scores for all 392 assessments, broken down in the three levels of FAIR characteristics: Essential, Intermediate, and Advanced. Each score is normalized to a maximum of 100 and does not take into account questions that are not applicable.

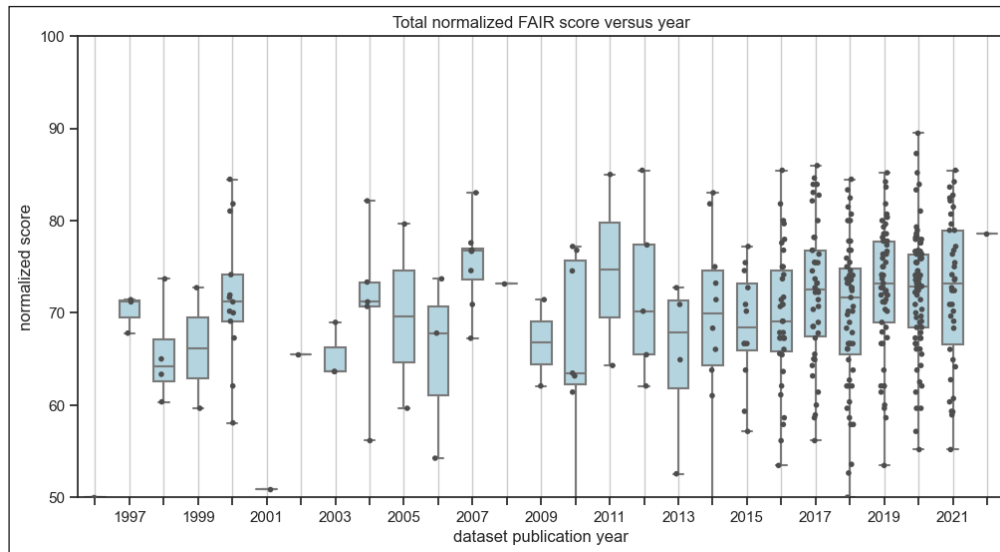
### SCORES BY PUBLICATION YEAR

The publication years of the dataset assessments range from 1987 to 2022 (Figure 4). Datasets were selected in order to examine pre-FSP policy and post-FSP policy scores (see the section ‘Selection Method’ for a description of these two groups).



**Figure 4** Bar chart showing the count distribution of the 392 datasets by publication year.

For the overall dataset of FAIR rubric assessments, there is not a discernable increase in FAIR scores over time (Figure 5).



**Figure 5** Box plots and data points showing the total normalized Findable, Accessible, Interoperable, and Reusable (FAIR) scores for datasets by publication year.

We also looked at scores by science discipline and hosting repository. The results for science discipline did not show significant trends and are not included here. Results by hosting repository were presented and discussed within the USGS. While important from an internal management viewpoint, a direct comparison between USGS repositories was not the emphasis of this paper, and we decided to publicly share results only on the collection as a whole.

### FAIR CRITERIA MOST OFTEN PRESENT

#### Questions with the highest number of ‘Yes’ answers

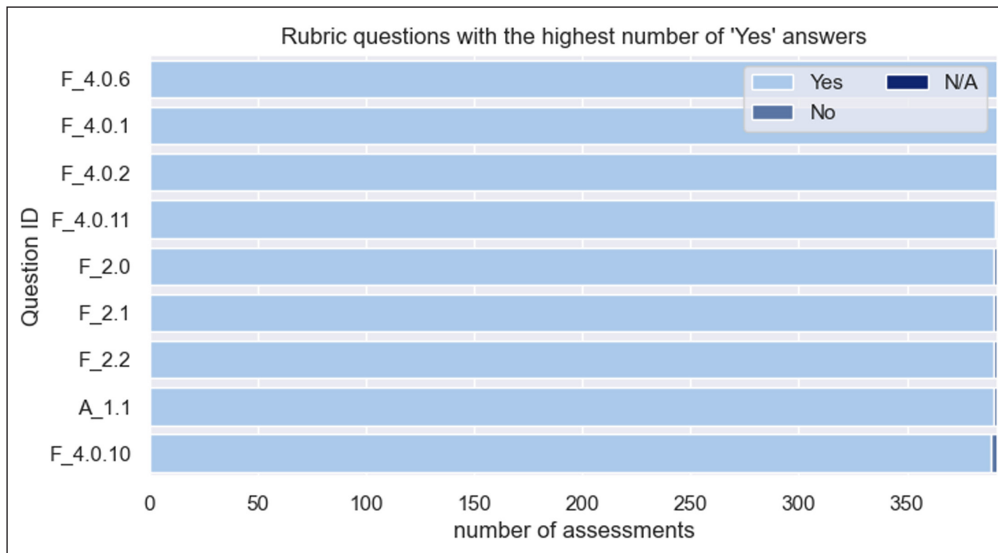
Ten questions had either 100% or 99% of the datasets meeting the criteria (Table 2 and Figure 6).

Three rubric questions had 100% of datasets meeting the criteria. For these questions, all datasets received ‘Yes’ answers. The SDC requires the specific metadata fields that are addressed by these three questions, so it is to be expected that all datasets within the sample set met the criteria. Seven additional rubric questions had greater than 99% of datasets meeting the criteria.

QUESTION ID	NUMBER YES	QUESTION
F_4.0.6	392	Is the following descriptive information included in the data release’s metadata? <b>Data publication date</b>
F_4.0.1	392	Is the following descriptive information included in the data release’s metadata? <b>Title</b>
F_4.0.2	392	Is the following descriptive information included in the data release’s metadata? <b>Description (e.g., Abstract, Summary, Purpose)</b>
F_4.0.11	391	Is the following descriptive information included in the data release’s metadata? <b>Keywords</b>
F_2.0	390	Is a separate identifier assigned for the data release’s metadata record?
F_2.1	390	Is the assigned identifier persistent?
F_2.2	390	Is the assigned identifier unique (i.e., has a unique value)?
A_1.1	390	Is this landing page publicly accessible?
F_4.0.10	389	Is the following descriptive information included in the data release’s metadata? <b>Temporal information</b> associated with the data release (e.g., start date and end date for when data were collected)
F_4.0.9	388	Is the following descriptive information included in the data release’s metadata? <b>Geographic location(s)</b> associated with the data release (e.g., coordinates)

**Table 2** The 10 rubric questions with the highest number of ‘Yes’ answers.





**Figure 6** Stacked bar chart showing the distribution of 'Yes', 'No', and 'Not Applicable' (N/A) answers for the rubric questions with the highest number of 'Yes' answers.

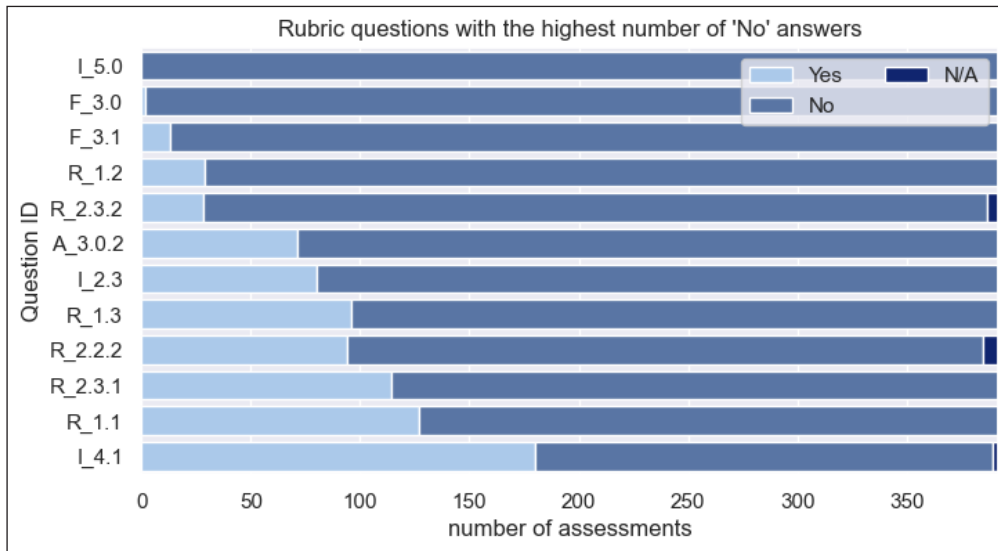
### FAIR CRITERIA LEAST OFTEN MET

#### Questions with the highest number of 'No' answers

There were 12 questions that had more 'No' than 'Yes' answers (Table 3 and Figure 7).

QUESTION ID	NUMBER NO	QUESTION
I_5.0	392	Is the data release described using <b>Resource Description Format (RDF) linked data with community-recognized ontologies</b> ?
F_3.0	390	Are the authors/originators' <b>ORCID identifiers viewable (to humans)</b> on the data release's landing page?
F_3.1	379	Are the authors/originators' <b>ORCID identifiers</b> provided in the data release's metadata?
R_1.2	363	Is the approved <b>USGS disclaimer statement</b> present on the data release's landing page?
R_2.3.2	359	Is the following information included with the data release's metadata? <b>Citation(s) to the citable (community recognized) guidelines or standards</b> used to describe the <b>data quality information</b> (e.g., using ISO 19157)
A_3.0.2	321	Is the following information included with the data release's landing page? <b>Data distributor contact information</b>
I_2.3	312	Are all data files in a format that is: <b>Available in multiple file formats</b>
R_1.3	296	Are <b>recommended reuses</b> present on the data release's landing page? AND/OR Are <b>known reuse limits</b> included on the data release's landing page?
R_2.2.2	291	Is the following information included with the data release's metadata? <b>Citation(s) to the citable (community recognized) guidelines or standards</b> used to describe the <b>process/methodology information</b>
R_2.3.1	278	Is the following information included with the data release's metadata? <b>Detailed data quality information</b> (e.g., data quality procedure documentation; data quality monitoring criteria during data collection, whether/how the completeness of the data files and their data values was evaluated)
R_1.1	265	Are <b>recommended reuses</b> included in the data release's metadata? AND/OR Are known reuse limits included in the data release's metadata?
I_4.1	209	Is information about <b>data value consistency</b> documented in the metadata?

**Table 3** The 12 rubric questions with the highest number of 'No' answers. ORCID: Open Researcher and Contributor IDs.



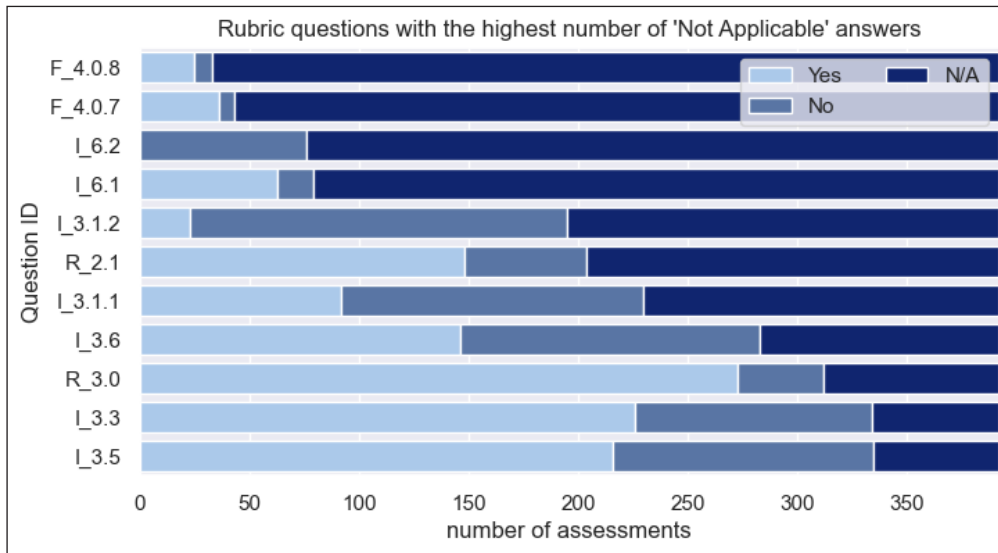
**Figure 7** Stacked bar chart showing the distribution of 'Yes', 'No', and 'Not Applicable' (N/A) answers for the rubric questions with the highest number of 'No' answers.

### Questions with the highest number of 'N/A' or 'Not Applicable' answers

Examining the number of 'Not Applicable' answers is important for understanding the data (Table 4 and Figure 8). A higher number of N/A answers for a rubric question means a lower number of yes/no data points that can be used to draw conclusions about the results. For questions where most of the answers are N/A, the FAIR score should be considered less reliable as a contributing factor in assessing the FAIRness of a USGS data release.

QUESTION ID	NUMBER 'NOT APPLICABLE'	
F_4.0.8	359	Is the following descriptive information included in the data release's metadata? If applicable, <b>data revision dates</b>
F_4.0.7	349	Is the following descriptive information included in the data release's metadata? If applicable, <b>data version</b>
I_6.2	316	If there are related data releases (other than source input datasets), are the relationships between the data releases: <b>Described using Resource Description Format (RDF)/linked data</b>
I_6.1	313	If there are <b>related data releases</b> (other than source input datasets), are the relationships between the data releases: Documented in the metadata
I_3.1.2	197	Does the data release's metadata contain the following information about the data release's attributes? <b>ALL names/labels are using citable and publicly available sources</b>
R_2.1	188	Is the following information included with the data release's metadata? If input datasets are used, the <b>citations to the input datasets</b>
I_3.1.1	162	Does the data release's metadata contain the following information about the data release's attributes? <b>At least one name/label is using a citable and publicly available source</b>
I_3.6	109	Does the data release's metadata contain the following information about the data release's attributes? <b>Allowable data values</b>
R_3.0	80	<b>Related resources</b> documented in the data release's metadata (e.g., project website, publications, use cases, job aids, user's guide, data processing code with readme, product algorithm document)
I_3.3	58	Does the data release's metadata contain the following information about the data release's attributes? <b>Units</b>
I_3.5	57	Does the data release's metadata contain the following information about the data release's attributes? <b>Data value range</b>

**Table 4** The 11 rubric questions with the highest number of 'Not Applicable' answers.



**Figure 8** Stacked bar chart showing the distribution of 'Yes,' 'No,' and 'Not Applicable' (N/A) answers for the rubric questions with the highest number of 'Not Applicable' answers.

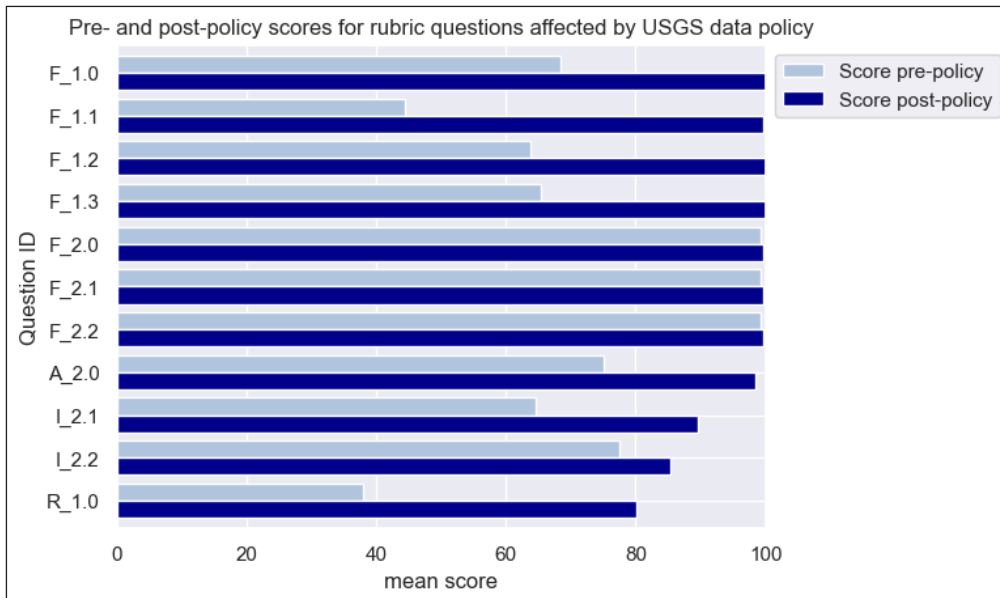
### Questions with the most improvement

#### Comparison of subsets published before and after data policy implementation

We compared the datasets published before the 2016 data policies with those published after. The following list is sorted based on the measure of the change between the two groups (Table 5).

QUESTION ID	CHANGE POST-POLICY	SCORE PRE-POLICY	SCORE POST-POLICY	
F_1.1	55.3	44.4	99.6	Is the assigned <b>identifier persistent</b> ?
R_1.0	42.3	37.9	80.2	Is an approved USGS <b>disclaimer statement</b> included in the data release's metadata?
A_4.1	40.7	54.8	95.5	Can users obtain the data release's <b>metadata files</b> by manual actions (human)
F_1.2	36.3	63.7	100	Is the assigned <b>identifier unique</b> (i.e., has a unique value)?
F_1.3	34.7	65.3	100	Is the assigned <b>identifier viewable</b> on the data release's <b>landing page</b> ?
F_1.0	31.5	68.5	100	Is an <b>identifier</b> assigned for the data release and documented in the data release's <b>metadata record</b> ?
I_2.1	25	64.5	89.6	Are all data files in a <b>format that is: Non-proprietary</b> (open format, i.e., accessible via free software)
A_2.0	23.5	75	98.5	Does the data release's <b>identifier resolve</b> to the human readable <b>landing page</b> ?
I_6.1	15.7	70	85.7	If there are related data releases (other than source input datasets), are the <b>relationships between the data releases: Documented</b> in the metadata
I_2.4	15.1	83.1	98.1	Are all data files in a <b>format that is: Expected</b> or commonly used by the relevant research community
R_2.2.2	14.1	14.8	28.9	Is the following information included with the data release's metadata? Citation(s) to the citable (community recognized) guidelines or standards used to describe the <b>process/methodology information</b>
I_3.3	10.2	60.4	70.6	Does the data release's metadata contain the following information about the data release's attributes? <b>Units</b>

**Table 5** The 12 rubric questions with the highest increase in 'Yes' questions after USGS data policy was implemented in 2016.



**Figure 9** Horizontal bar plot showing the 11 questions that address elements affected by the USGS data policy implementation in 2016, showing an increase in number of ‘Yes’ answers for all questions.

Eleven of the 62 questions address elements that are affected by the 2016 data policies (Figure 9). For example, F\_1.1, F\_1.2, and F\_1.3 ask about the identifier assigned to the data. USGS policy now requires digital object identifiers (DOIs) for all data release products, and, as shown in Figure 9, the relevant scores increased.

## DISCUSSION

### ANALYSIS OF RESULTS

At the outset of this project, we expected that the overall scores for Findable and Accessible would be relatively high, due to the existing FSP guidance and best practices followed by many USGS repositories and data authors. The study results confirmed this (Figure 2). All but one of the rubric questions scoring 99–100% positive came from the Findable category; metadata elements that underpin this category are largely bibliographic and are generally less difficult to complete. We expected scoring for Interoperable and Reusable to be lower, as the metadata sections that support these categories are generally more nuanced, labor intensive, and subject to inconsistencies in interpretation by metadata authors. This was also observed in the study results.

Each question in the rubric was categorized into Essential, Intermediate, or Advanced, based on its perceived level of importance to the key characteristics of FAIR. The datasets performed relatively well on elements most vital for FAIR (Figure 3), which are often policy requirements for public release. The more advanced characteristics were less commonly employed, likely because they have more obscure applicability, with limited to no guidance or precedent in USGS data releases.

We expected FAIR scores to increase over time, especially for the periods before and after the implementation of the 2016 USGS data policies. The data release process has become more standardized since the release of those policies, and it is better supported by USGS tools and resources, such as those created for the ScienceBase data release process (Hutchison et al. 2021). While we did not observe significant changes in *total* FAIR scores over time within our sample set (Figure 5), a closer look at individual questions revealed interesting trends. Questions affected by policy, most of which are categorized in the rubric as ‘Essential,’ showed significant improvement in the periods before and after the data policies (Table 5 and Figure 9). These included questions relating to persistent identifiers, approved disclaimer statements, data file formats, and file accessibility.

### NOTES ON LIMITATIONS DUE TO THE SCOPE OF THE PROJECT

USGS repositories differ in the technical capabilities that they offer to end users; for example, some provide programmatic access to data or checksums for data integrity. As described in the section ‘Rubric creation,’ we separated out the repository-dependent FAIR characteristics from the core set of questions in the rubric. Although this study focuses on the characteristics that data authors can control, there are additional actions that repositories and bureau-level managers can take to further improve FAIRness. These actions and considerations are listed

within the rubric (Hutchison et al. 2023) and can be used as a basis for future discussions about repository capabilities within the USGS.

Data maturity and analysis readiness are important concepts that are not included in this analysis. We decided to focus on the FAIR principles instead of the broader concepts of analysis-readiness and data maturity. This decision was based on the need to evaluate a wide variety of datasets in a consistent and quantitative way. The USGS produces data products that are diverse in terms of data type, product complexity, and scientific discipline (e.g., imagery data, time series data, tabular data, geospatial data). A goal of this project was to quantify certain qualitative characteristics across datasets. To accomplish this, the new rubric needed to be broadly applicable, so that resulting metrics could be aggregated and analyzed. The concept of analysis readiness is broad and depends on variables such as data type, file format, subject matter, and intended use of the data (e.g., criteria for one file format may not apply to other formats, and a published dataset may be ready for ingest and use by one type of application but not others). The team decided that focusing on the FAIR principles was the most practical approach given our task and time frame. Future studies could focus more closely on specific data types or domains to provide useful guidance for optimizing their utility.

For notes on additional limitations of the study, see the ‘Sample set composition’ section of this paper.

## STRATEGIES FOR BETTER ALIGNMENT WITH FAIR PRINCIPLES

Our analysis resulted in strategies organized around four key areas: USGS data repository characteristics; training and communities of practice; data management policy considerations; and metadata standards, tools, and best practices (Table 6). The column ‘FAIR Workshop proposed activity’ shows the strategies from Lightsom et al. (2022) that align with ours.

	STRATEGY	CATEGORY	FAIR WORKSHOP PROPOSED ACTIVITY	FAIR ELEMENT IMPROVED	LEVEL OF EFFORT	ROI
R1	Convene repository managers to develop core shared standards for presentation of/access to data and metadata via landing pages	Data Repositories	5-1 5-12	F,A	M	M
R2	Move repositories towards standard processes, workflows, and services for intake of new data releases	Data Repositories	5-5 5-17 5-21	F,A	M	H
P1	Reevaluate minimum characteristics for repositories to be considered for inclusion in the acceptable repositories list	Policy	5-1	F,A	M	M
P2	Clarify requirements for and implementation of disclaimers, licenses, and constraints on use and access	Policy	2-1 2-2 2-14	A,R	M	M
P3	Institute peer review process for comprehensive data management plans at project outset	Policy	7-2	A,R	M	H
C1	Convene working group to improve data quality documentation practices in metadata	Community & Training	-	R	H	H
C2	Use community-based approach to define data dictionaries that support linked open data	Community & Training	3-6	I	H	H
C3	Convene repository managers to develop consistent practices for documenting version history and links between versions	Community & Training	7-3	F,A	M	M
C4	Consider developing training program for writing data management plans that anticipate and plan for FAIR requirements	Community & Training	7-2	F,A,R	M	H

**Table 6** Strategies.

Table legend: L: low, M: medium, H: high, ROI: Return on investment. FAIR (Findable, Accessible, Interoperable, and Reusable) workshop proposed activities: the numbers reference proposed activities in Lightsom et al. (2022).

	STRATEGY	CATEGORY	FAIR WORKSHOP PROPOSED ACTIVITY	FAIR ELEMENT IMPROVED	LEVEL OF EFFORT	ROI
C5	Use community-based approach to evaluate open and machine-readable data formats and develop best practices for implementation by scientists and repositories	Community & Training	5-17 5-21	A,R	M	H
C6	Consider developing training to support broader understanding of persistent identifiers for access, credit, citation, and use of data	Community & Training	-	A,R	L	M
C7	Leverage community groups to support adoption of shared classification schemes and vocabularies to describe and characterize data assets	Community & Training	-	F,A,I	M	M
M1	Consider adoption of ISO to facilitate inclusion of more precise, unambiguous, and FAIR descriptions of dataset characteristics	Metadata	-	F,A,I,R	H	H
M2	Optimize metadata editor tools to document data in a standards-agnostic language, to facilitate interoperability with applications, standards, and workflows	Metadata	-	F,A,I,R	H	H
M3	Promote best practices for reusable metadata elements that are citable and discoverable on their own	Metadata	2-4 3-4 3-6	R	M	M
M4	Improve metadata tools, as informed by usability analyses	Metadata		F,A,I,R	M	M
M5	Evaluate opportunities to apply AI/ML tools to metadata assessments, possibly broadening range of applicability	Metadata		F,A,I,R	M	H

## REPOSITORIES

The USGS maintains multiple robust data repositories. Of the ten repositories represented in this study, seven are USGS assets.

Inconsistencies in what repositories choose to display for human readability on a data release's landing page impact the extent to which a data release can fully meet Findable and Accessible criteria. For example, only 18% of the data releases sampled include distribution contact information on the landing page, and only 7% sampled included a standard USGS distribution liability statement for public datasets. This information is available from the metadata but would require a user to open the metadata file from the landing page to access it.

**Strategy R1** asks that USGS repository managers develop a joint template or approach for the display of key metadata fields on landing pages. This aligns with the proposed activities of Lightsom et al. (2022), who call for 'human-readable indicators on landing pages,' particularly for quick identification of information that might not be included in the metadata. Examples of core landing page information that could be standardized across repositories include persistent identifiers for data and authors; distribution contact information; use constraints on or limitations of the data; and a full listing of downloadable files in the data release.

**Strategy R2** is for USGS repository management teams to better standardize their ingest processes, workflows, and services for new data releases. This could help with consistency in display of author ORCID's (Open Researcher and Contributor IDs), version history, and revision details of a dataset. These findings align with proposed activities from Lightsom et al. (2022) for 'standardized curation methods in [USGS] repositories' and the need for appropriate 'funding and staffing to ensure that USGS repositories...are trusted, reliable, curated, and efficient.'

Because of the variability between different data repositories, systems, and catalogs that are used for USGS data releases, it would be beneficial for a policy committee, working together with repository managers, to define the minimum criteria for designation as an approved USGS data repository (**Strategy P1**). There are two levels of approved status for USGS repositories: acceptable and trusted. USGS systems that have been designated as ‘acceptable repositories’ are generally considered to be mature; however, these systems could be encouraged to advance their capabilities and services to levels that would qualify them for USGS ‘trusted digital repository’ designation.

Policy committees could also address the inconsistent application and use of statements addressing licensing, access constraints, use constraints, and liability statements (**Strategy P2**). Licensing and access constraints are applied inconsistently in the sampled USGS dataset metadata. Many datasets do not specify a license, an outcome of the absence of a data license field in the CSDGM standard. For example, some USGS metadata authors attempt to address licensing in a narrative statement about access constraints, and will indicate a license statement, such as ‘Public Domain’ or ‘Creative Commons CCO,’ to indicate that the federally produced data product is freely available for unrestricted use.

We believe that moving from CSDGM towards the ISO 19115 suite of metadata standards, which explicitly define and distinguish between licensing of data and constraints upon use, will help address these challenges (see ‘Metadata’ section below).

Determining and specifying licensing, liability, constraints, persistent identifiers, and other fields often can be best addressed in the planning stage of a research project. Data management plans were officially required for all new research projects upon institution of the FSP data policies in 2016. However, decentralized oversight of data management planning and data agreements with non-USGS partners has led to inconsistency in format, extent, and formal review of data management plans. **Strategy P3** is to develop minimum required elements for data management plans that would ensure thoughtful consideration of FAIR characteristics of the data at the outset of the research project.

These strategies align with several activities proposed by Lightsom et al. (2022), including establishment of policies for machine-readable licenses, approved and standardized language for constraints and disclaimers, and development of data management plans that will ensure findability of data.

## COMMUNITY AND TRAINING

This section describes actions that can be applied through leveraging communities of practice and training programs. There are seven community and training strategies; the first two are discussed here.

**Strategy C1** is for a working group to focus on improving data quality documentation practices, which could include defining minimum criteria for metadata fields and developing training to support best practices. Study results show a trend of inconsistent or incomplete documentation of data quality in our sample set. Some of the data quality results did not show an improvement over time, and some even showed a decline when we compared the subsets published before and after the data policies.

We hypothesize that this could be related to the dramatic increase in the number of data releases published following the new policy requirements. More metadata records are now being created by data authors who are newer to metadata creation and do not necessarily have dedicated time to spend on the task. Pre-policy, metadata records were more often written by a smaller, highly trained cohort of data managers working within those programs and centers with a long history of documenting and releasing data.

In the future, guidance and training could be specialized for both data authors, who have the most in-depth knowledge of their data quality processes, and data managers, who are often the dedicated metadata creators at their science centers.

**Strategy C2** is to improve the interoperability of USGS data through the creation and use of enterprise and community of practice data dictionaries. USGS does not have an enterprise data dictionary, and except for a few large, real-time data systems, we have few data dictionaries

available within our major scientific domains. The result is that each of our datasets essentially has its own unique data dictionary, and efforts to assess interoperability are difficult.

While the percentage of assessments with data attribute labels and definitions in the metadata was relatively high (82% and 79% ‘Yes’, respectively), other important attribute descriptive elements were included less often, including the units of measure (65% ‘Yes’), the data value ranges (65%), the allowed data values (52%), and the means for assessing attribute value accuracy (55%) and consistency (46%) in the data values collected. The absence of these details in the metadata not only inhibits interoperability, even by manual means, but also introduces challenges for users in evaluating the fitness of the data for a particular use.

Many of the FAIR Workshop proposed activities call for teams to organize or create content to improve FAIRness, such as ‘create a team to develop or discover standard data dictionaries and provide them online to encourage their use and enable citation in metadata.’

A strong collaboration ethic in USGS communities of practice has supported activities addressing all components of the data lifecycle. The USGS Community for Data Integration (CDI) and its many working groups have been instrumental in bringing together researchers, data managers, policy experts, and IT specialists to identify and address challenges associated with data management and integration (Hsu et al. 2022; USGS 2023). We envision the CDI playing an important role in enacting some of the strategies in this section.

## METADATA

There are five strategies described in this section; the first two are discussed below.

One consistent theme that emerged from this study is that the FAIRness of our data releases is limited in part by continued investment in an older metadata content standard – CSDGM – which has not been updated in more than two decades, and which does not readily support key FAIR principles such as the application and use of persistent identifiers and the ability to document and easily link together hierarchical and associative metadata records. For example, CSDGM does not include capabilities to uniquely associate authors with ORCIDs, organizations with identifiers such as ROR IDs (Research Organization Registry IDs), or vocabulary concepts with Uniform Resource Identifiers (URIs). While many of these identifiers can be inserted in text fields in the metadata, there are no semantic operators that enable machines to identify them as a type of unique persistent identifier, and to parse them accordingly.

**Strategy M1** is to move to more modern documentation standards with supported metadata creation tools that provide strong usability, employ governance of concept domains and controlled vocabularies, fully leverage persistent identifiers, and facilitate capabilities to detail relationships between and among data assets. **Strategy M2** describes an approach to facilitate this move: leveraging a metadata tool that uses a standards-agnostic language to enable interoperability with metadata standards, profiles, and workflows that meet USGS and Federal requirements. The USGS and other Department of the Interior bureaus are actively developing the mdEditor suite of tools (USGS and FWS n.d.) to accomplish these goals.

These strategies align with a key activity proposed by Lightsom et al. (2022) for ‘machine-actionable metadata [to] improve data discovery and reuse.’

## ADDITIONAL STRATEGIES

Feedback from colleagues and partners describes the USGS FAIR rubric as labor intensive in its application. It requires a human reviewer to examine the metadata and the repository landing page of the dataset, determine a ‘Yes’ or ‘No’ response to each question, and manually enter that response in the spreadsheet. The common request we receive from users is to transform the rubric spreadsheet into an automated online tool that can evaluate and score each question without the need for significant human input. Projects that have successfully applied automated scoring to FAIR evaluations include Devaraju and Hurt (2021), Clarke et al. (2019), and Jones et al. (2019).

Automation would be relatively straightforward for 22 of the 62 questions in the rubric, as they are scored based on the presence or absence of content within the metadata or landing page, or by adherence to expected format or patterns (e.g., a valid URI or identifier). Automation of the rubric would be more challenging for the questions where content needs to be evaluated to check that it is complete, logically organized, and adherent to policy (e.g., data processing steps,



access constraints, or accuracy checks). A next step would be to explore machine learning and artificial intelligence applications to determine the feasibility of an automated tool that could be developed and then trained with exemplar data releases to recognize and appropriately score data releases.

Additionally, we recognized that lower scores for certain questions from the rubric were a consequence of broader issues, including improper interpretation of USGS policies, or incomplete or inadequate data management planning at earlier stages of the data collection effort. In many cases, the sampled data release might have performed better on certain FAIR factors if earlier consideration had been made about topics such as repository selection, legal issues surrounding access constraints, documentation of calibration and accuracy of tools and measurements, and use of existing data dictionaries to characterize certain data parameters. A next step to be considered would be to decompose the individual questions in the rubric and align them to the USGS Data Management Lifecycle (Faundeen et al. 2014), which could support the collection of important characteristics of FAIR data at the most appropriate time in the data lifecycle.

Finally, a follow-up study, or series of studies, could measure the progress towards more FAIR data. Such studies could take many forms, including:

- In the event of revisions to data releases sampled in the original study, re-analyzing those releases using the rubric to determine whether the revisions have improved their FAIR scores;
- expanding the number of data releases analyzed to further contextualize and confirm our standing with respect to FAIR;
- collaborating with individual USGS repositories to perform a deeper evaluation of holdings and to learn whether changes in repository practices and policies could improve FAIR scores; and
- working with data managers from individual USGS science centers and programs to perform FAIR analyses of their existing data releases and uncover trends and pathways towards improved FAIR scores.

## CONCLUSION

The primary goal of the State of the Data Project was to develop and implement a methodology to assess the FAIRness of published USGS data products. Based on the results, we generated possible strategies for improving alignment of USGS data with the FAIR principles.

This list of strategies is designed to be targeted, practical, and an efficient use of resources. We believe that coordination across USGS repositories and support for migration to a more current metadata standard would provide a wide range of benefits and improve FAIRness in many of the elements we assessed. There is significant overlap between this project's suggested strategies and those written by the FAIR Workshop team (Lightsom et al. 2022). We interpret this to mean that the two different approaches, one based on qualitative discussions and one on quantitative assessments, can both be effective ways to understand the state of USGS data and identify areas for improvement.

An important takeaway for us is that a quantitative FAIR assessment across a diverse range of data products is feasible and can provide useful insights. The fact that we focused exclusively on USGS data was helpful: the USGS publishes data products that have metadata in a standardized format and that have passed a rigorous review process. We benefited from a dedicated team of data managers who contributed significant time and effort to calibrate and document scoring guidance. Nevertheless, the limitations of our time frame and the diversity of the datasets necessitated a manual evaluation based largely on presence or absence of content. For future studies, we envision new opportunities for automation based on recent advances in artificial intelligence and natural language processing.

## DATA ACCESSIBILITY STATEMENT

Data and Metadata are openly available at Hutchison, VB., Zolly, LS., Norkin, T, Hsu, L and Hou, C-Y. 2023. USGS State of the Data Project: Rubric and Assessment Data. US Geological Survey data release. DOI: <https://doi.org/10.5066/P97V4XA4>.

The additional files for this article can be found as follows:

- **Supplemental File 1: FAIR Rubric Questions.** A list of all the questions in the USGS FAIR Rubric (Hutchison et al. 2023). 'FAIR': Findable, Accessible, Interoperable, Reusable. DOI: <https://doi.org/10.5334/dsj-2024-022.s1>
- **Supplemental File 2: Data Presented in Figures 1–4 and Tables 2–5.** Data from the figures in this paper shared in tabular format. For the complete dataset, see Hutchison et al. 2023. DOI: <https://doi.org/10.5334/dsj-2024-022.s2>
- **Supplemental File 3: Example Datasets.** A table of example datasets that were evaluated using the USGS FAIR Rubric. The table includes the FAIR score assigned for each dataset. These were the five highest scores in the sample set. DOI: <https://doi.org/10.5334/dsj-2024-022.s3>

## ACRONYMS IN ADDITIONAL FILES

CSDGM: Content Standard for Digital Geospatial Metadata

DOI: Digital Object Identifier

FAIR: Findable, Accessible, Interoperable, Reusable

FGDC: Federal Geographic Data Committee

ISO: International Organization for Standardization

ORCID: Open Researcher and Contributor IDs

USGS: US Geological Survey

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the US government.

## ACKNOWLEDGEMENTS

Special acknowledgement: Chung-Yi (Sophie) Hou (ORCID [0000-0002-8087-1775](https://orcid.org/0000-0002-8087-1775)) completed phase 1 of this study and was integral to the creation of the USGS FAIR Rubric, the assessment's methodology, and the data assessment process.

Data assessments were completed by: Grace C. Donovan, Chung-Yi Hou, Amanda N. Liford, Madison L. Langseth, Ricardo McClees-Funinan, Brittany G. Waltemate

The authors would like to thank the following people at the USGS for their help in shaping the rubric:

Matt Cannister, Susie Cochran, VeeAnn Cross, Katherine Dahm, Linda Debrewer, Grace Donovan, Ricardo McClees-Funinan, Arnell Forde, Madison Langseth, Amanda Liford, Ryan Longhenry, Tim Mentele, John Reed, Peter Schweitzer, Brittany Waltemate, and Dennis Walworth.

The authors would also like to thank Michaela Johnson (USGS) and anonymous external reviewers for their suggestions for improving this paper.





Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the US Government.

## COMPETING INTERESTS

The authors have no competing interests to declare.

## AUTHOR CONTRIBUTIONS

VH: conceptualization, writing (original draft, review, editing), supervision; TN: conceptualization, methodology, data curation, writing (original draft, review, editing); LZ: conceptualization, writing (original draft, review, editing); LH: analysis, visualization, writing (original draft, review, editing).

- Vivian B. Hutchison**  [orcid.org/0000-0001-5301-3698](https://orcid.org/0000-0001-5301-3698)  
US Geological Survey, Science Analytics and Synthesis, Denver, Colorado, USA
- Tamar Norkin**  [orcid.org/0000-0003-0797-3940](https://orcid.org/0000-0003-0797-3940)  
US Geological Survey, Science Analytics and Synthesis, Denver, Colorado, USA
- Lisa S. Zolly**  [orcid.org/0000-0003-3595-7809](https://orcid.org/0000-0003-3595-7809)  
US Geological Survey, Science Analytics and Synthesis, Denver, Colorado, USA
- Leslie Hsu**  [orcid.org/0000-0002-5353-807X](https://orcid.org/0000-0002-5353-807X)  
US Geological Survey, Science Analytics and Synthesis, Denver, Colorado, USA

## REFERENCES

- Clarke, DJB, Wang, L, Jones, A, Wojciechowicz, ML, Torre, D, Jagodnik, KM, Jenkins, SL, McQuilton, P, Flamholz, Z, Silverstein, MC, Schilder, BM, Robasky, K, Castillo, C, Idaszak, R, Ahalt, SC, Williams, J, Schurer, S, Cooper, DJ, de Miranda Azevedo, R, Klenk, JA, Haendel, MA, Nedzel, J, Avillach, P, Shimoyama, ME, Harris, RM, Gamble, M, Poten, R, Charbonneau, AL, Larkin, J, Brown, CT, Bonazzi, VR, Dumontier, MJ, Sansone, S-A and Ma'ayan, A** 2019. FAIRshake: Toolkit to evaluate the FAIRness of research digital resources. *Cell Systems*, 9(5): 417–421. DOI: <https://doi.org/10.1016/j.cels.2019.09.011>
- Devaraju, A and Huber, R** 2021. An automated solution for measuring the progress toward FAIR research data. *Patterns*, 2(11): 100370. DOI: <https://doi.org/10.1016/j.patter.2021.100370>
- FAIRsFAIR** 2022. FAIRsFAIR: Fostering FAIR data practices in Europe. Available at <https://www.fairsfair.eu> [Last accessed June 01, 2023].
- Faundeen, J, Burley, TE, Carlino, JA, Govoni, DL, Henkel, HS, Holl, SL, Hutchison, VB, Martin, E, Montgomery, ET, Ladino, C, Tessler, S and Zolly, LS** 2014. The United States Geological Survey Science Data Lifecycle Model. US Geological Survey Open-File Report 2013–1265. DOI: <https://doi.org/10.3133/ofr20131265>
- Federal Geographic Data Committee (FGDC)** 1998. FGDC-STD-001–1998. Content standard for digital geospatial metadata, Version 2. Available at <https://www.fgdc.gov/standards/projects/metadata/base-metadata> [Last accessed October 01, 2023].
- Fundamental Science Practices Advisory Committee (FSPAC)** 2011. US Geological Survey Fundamental Science Practices. US Geological Survey Circular 1367. DOI: <https://doi.org/10.3133/cir1367>
- Fundamental Science Practices Advisory Committee (FSPAC)** 2023. Update on US Geological Survey Fundamental Science Practices. US Geological Survey Circular 1503. DOI: <https://doi.org/10.3133/cir1503>
- Go FAIR** n.d. Go FAIR Initiative. Available at <https://www.go-fair.org/go-fair-initiative> [Last accessed 2020].
- Habermann, T and Jones, MB** 2020. Data Observation Network for Earth (DataONE)/Metadata Game Changers FAIR Metadata Recommendations. Unpublished version. Presented at <https://2020esipwintermeeting.sched.com/event/VaXT/fair-metadata-recommendations> [Last accessed March 25, 2020].
- Hsu, L, Liford, AN and Donovan, GC** 2022. Community for data integration 2020 annual report. U.S. Geological Survey Open-File Report 2022–1034. DOI: <https://doi.org/10.3133/ofr20221034>
- Hutchison, VB, Norkin, T, Langseth, ML, Ignizio, DA, Zolly, LS, McClees-Funinan, R and Liford, A** 2021. Leveraging existing technology: Developing a trusted digital repository for the U.S. Geological Survey. *International Journal of Digital Curation*, 16(1): 23–23. DOI: <https://doi.org/10.2218/ijdc.v16i1.741>
- Hutchison, VB, Zolly, LS, Norkin, T, Hsu, L and Hou, C-Y** 2023. USGS State of the Data Project: Rubric and Assessment Data. US Geological Survey data release. DOI: <https://doi.org/10.5066/P97V4XA4>
- Jones, MB, Slaughter, P, Habermann, T and Gordon, S** 2019. metadig-checks: MetaDIG suites and checks for data and metadata improvement and guidance. Available at <https://github.com/NCEAS/metadig-checks> [Last accessed 2020].
- Lightsom, FL, Hutchison, VB, Bishop, B, Debrewer, LM, Latysh, N and Stall, S** 2022. Opportunities to improve alignment with the FAIR Principles for US Geological Survey data. US Geological Survey Open-File Report 2022–1043. DOI: <https://doi.org/10.3133/ofr20221043>
- Office of Management and Budget (OMB)** 2013. Open Data Policy: Managing Information as an Asset (OMB Memorandum M-13-13). Available at <https://digital.gov/resources/open-data-policy-m-13-13>.
- OSTP** 2013. Memorandum for the heads of executive departments and agencies: Increasing access to the results of federally funded scientific research. Available at <https://www.whitehouse.gov/wp-content/uploads/2022/08/08-2022-OSTP-Public-Access-Memo.pdf>.
- OSTP** 2022. Memorandum for the heads of executive departments and agencies: Ensuring free, immediate, and equitable access to federally funded research. DOI: <https://doi.org/10.21949/1528361>
- Peng, G** 2023. Finding harmony in FAIRness. *Eos*, 104. DOI: <https://doi.org/10.1029/2023EO230216>
- Peng, G** n.d. Data Stewardship Maturity Matrix (DSMM) resources. Available at <https://ncics.org/portfolio/data-stewardship/dsmm/> [Last accessed 2021].

- Peng, G, Downs, RR, Ramapriyan, HK, Parsons, MA, Moroni, DF, Liu, Z, Khalsa, SJS, Mears, C, Wei, Y, Ramachandran, B, Smith, S and NASA O'FAIR Working Group** 2023. An overview of community FAIR practices – NASA O'FAIR WG inception report. Document ID: NASA-OFAIR-ESDSWG-DOC-0001. DOI: <https://doi.org/10.5067/DOC/ESCO/ESDSWG-0001V1>
- Peng, G, Privette, JL, Kearns, EJ, Ritchey, NA and Ansari, S** 2015. A unified framework for measuring stewardship practices applied to digital environmental datasets. *Data Science Journal*, 13(0): 231–252. DOI: <https://doi.org/10.2481/dsj.14-049>
- Ramachandran, R, Bugbee, K and Murphy, K** 2021. From open data to open science. *Earth and Space Science*, 8(5): e2020EA001562. DOI: <https://doi.org/10.1029/2020EA001562>
- RDA FAIR Data Maturity Model Working Group** 2020. FAIR Data Maturity Model: Specification and guidelines (1.0). DOI: <https://doi.org/10.15497/RDA00050>
- US Congress** 2017. The Open, Public, Electronic, and Necessary Government Data Act or the OPEN Government Data Act (H.R. 1770).
- US Congress** 2018. Foundations for Evidence-Based Policymaking Act of 2018 (H.R. 4174).
- US Geological Survey (USGS)** 2016. Public access to results of federally funded research at the US Geological Survey. Available at <https://www.usgs.gov/office-of-science-quality-and-integrity/public-access-results-federally-funded-research-us> [Last accessed October 1, 2023].
- US Geological Survey (USGS)** 2017a. 502.6- Fundamental science practices: Scientific data management. Available at <https://www.usgs.gov/survey-manual/5026-fundamental-science-practices-scientific-data-management>.
- US Geological Survey (USGS)** 2017b. 502.7- Fundamental science practices: Metadata for USGS scientific information products including data. Available at <https://www.usgs.gov/survey-manual/5027-fundamental-science-practices-metadata-usgs-scientific-information-products>.
- US Geological Survey (USGS)** 2017c. 502.8- Fundamental science practices: Review and approval of scientific data for release. Available at <https://www.usgs.gov/survey-manual/5028-fundamental-science-practices-review-and-approval-scientific-data-release>.
- US Geological Survey (USGS)** 2017d. 502.9- Fundamental science practices: Preservation requirements for digital scientific data. Available at <https://www.usgs.gov/survey-manual/5029-fundamental-science-practices-preservation-requirements-digital-scientific-data>.
- US Geological Survey (USGS)** 2023. CDI activities in FY 2023. Available at <https://www.usgs.gov/community-data-integration-activities/cdi-activities-fy-2023> [Accessed March 6, 2024].
- US Geological Survey (USGS)** n.d.-a. Data management – Data release. Available at <https://www.usgs.gov/data-management/data-release> [Last accessed October 1, 2023].
- US Geological Survey (USGS)** n.d.-b. USGS Science Data Catalog (SDC). Available at <https://data.usgs.gov/datacatalog> [Last accessed October 1, 2023].
- US Geological Survey (USGS) and US Fish and Wildlife Service (USFWS)** n.d. mdToolkit. Available at <https://www.mdtoolkit.org> [Last accessed October 1, 2023].
- Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, J-W, da Silva Santos, LB, Bourne, PE, Bouwman, J, Brookes, AJ, Clark, T, Crosas, M, Dillo, I, Dumon, O, Edmunds, S, Evelo, CT, Finkers, R, Gonzalez-Beltran, A, Gray, AJG, Groth, P, Goble, C, Grethe, JS, Heringa, J, 't Hoen, PAC, Hooft, R, Kuhn, T, Kok, R, Kok, J, Lusher, SJ, Martone, ME, Mons, A, Packer, AL, Persson, B, Rocca-Serra, P, Roos, M, van Schaik, R, Sansone, S-A, Schultes, E, Sengstag, T, Slater, T, Strawn, G, Swertz, MA, Thompson, M, van der Lei, J, van Mulligen, E, Velterop, J, Waagmeester, A, Wittenburg, P, Wolstencroft, K, Zhao, J and Mons, B** 2016. The FAIR Guiding principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. DOI: <https://doi.org/10.1038/sdata.2016.18>
- Wilkinson, MD, Sansone, S-A, Schultes, E, Doorn, P, Bonino Da Silva Santos, LO and Dumontier, M** 2018. A design framework and exemplar metrics for FAIRness. *Scientific Data*, 5: 180118. DOI: <https://doi.org/10.1038/sdata.2018.118>

#### TO CITE THIS ARTICLE:

Hutchison, V B, Norkin, T, Zolly, L S and Hsu, L 2024 State of the Data: Assessing the FAIRness of US Geological Survey Data. *Data Science Journal*, 23: 22, pp. 1–20. DOI: <https://doi.org/10.5334/dsj-2024-022>

**Submitted:** 18 August 2023

**Accepted:** 23 March 2024

**Published:** 26 April 2024

#### COPYRIGHT:

© 2024 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.