JOURNAL OF POLITICAL STABILITY ARCHIVE

Online ISSN: 3006-5879 Print ISSN: 3006-5860
DOI: https://doi.org/10.63468/jpsa.4.2.06
Vol. 4 No. 2 (2026)
https://journalpsa.com.pk/index.php/JPSA/about

Recognized by: Higher Education Commission (HEC), Government of Pakistan

Meta Check: A Comprehensive Framework for Profiling and Repairing Metadata Quality in Open Data Repositories

Riaz Ahmed •
Bachelor's Student at Faculty of Higher IT School, National Research Tomsk State University, Tomsk, Russian Federation.
Riazahmed5231@gmail.com

Shakil Ahmed
Bachelor's Student at Department of Petroleum Engineering, Chattogram University of Engineering and Technology, Chattogram, Bangladesh.
Shakilahmed14580@gmail.com

Md Ruhul Ibna Khan Jesun
Bachelor's Student at TISP Molecular Engineering, National Research Tomsk State University, Russian Federation.
rikjesun@gmail.com

Shaik Abul Zaid
Bachelor's Student at Faculty of Higher IT School, National Research Tomsk State University, Tomsk, Russian Federation.
shaikabulzaid@outlook.com

•Corresponding Author

ABSTRACT

The quality of metadata that is pervasive in nature is critical in undermining the utility of open data repositories, which are the key component of the modern research and open government endeavours. Such shortcomings are very detrimental to data discovery, interoperability, and reuse, and, in effect, data silos form within a so-called interconnected ecosystem. The paper presents a new and holistic approach, named Meta Cheque, to profile the quality of metadata systematically, assess it, and fix it in different open data repositories. The high quality of our methodology lies in the fact that it is a four-tier process: (1) scalable harvesting of metadata of a large variety of portals, both general-purpose (e.g. Data.gov) and

domain-specific (e.g. the EDX of the National Energy Technology Laboratory), (2) multi-dimensional profiling of metadata quality based on quantifiable metrics of completeness, consistency, timeliness, and standard compliance (DCAT and FAIR), (3) the use of a suite of lightweight yet powerful automated repair methods, including ontology Findings of a large scale examination of more than 50,000 datasets suggest that there is a high level of metadata deficiency in general and domain-specific repositories. The use of our automated repair procedures resulted in significant quality improvements, and completeness scores were raised by a mean of 35 per cent, and consistency by more than 50. Search simulations conducted with federated search revealed a meaningful improvement in discoverability, whereby the average recall showed a 28 per cent improvement and the average precision showed a 22 per cent improvement. A more thorough case study (based on petroleum engineering data) found that domain-specific fixes, like matching terms to the NASA GCMD thesaurus, increased the accuracy of complex technical queries significantly, e.g., of concepts like "gamma ray log" and "3D seismic survey." We have definitively shown that metadata remediation can be done in a systematic, automated manner, which is not only technically feasible but is central to the development of an open data ecosystem that is actually integrated, functional, and trustworthy. This is more important in those data-intensive and big-stakes areas such as energy and geosciences, where the price of either not discovering or failing to interpret the data is extremely high.
Keywords: Metadata Quality, Open Data, Data Discovery, FAIR Principles, Data Profiling, Petroleum Engineering, Mining Engineering.

INTRODUCTION

The past ten years have seen an unparalleled explosion of open data repositories due to movements of open science, open governance, and open innovation around the world. In the United States, the Data.gov project and in the United Kingdom, the Data.gov.uk project have published large amounts of data in open access, as have institutional and general-purpose archives such as Zenodo and Figshare. These repositories are seen as the backbone of the new research, policy-making, and economic development, where data can be accessed, combined, and analysed freely to produce new knowledge.

Nevertheless, this vision is not quite fulfilled. One of such bottlenecks is the bad quality of metadata, which describes these datasets, and this problem is frequently underestimated. Metadata, which has been described as data about data, is what gives data the necessary sense so as to be interpretable, discoverable, and reusable. It contains details like the titles, description, authors, keywords, licence, time, and spatial coverage. In the absence of good metadata, a dataset is the book in the library without a title on its spine, a list of authors on the spine, or a book in the card catalogue; its presence may be known, but it has been practically forgotten.

The main issue is the inaccurate, partial, and non-standardised character of metadata between and within repositories. The contributors of the data, government

agencies, or individual researchers, usually do not have the time, knowledge, or motivation to offer rich standardised metadata. The outcome is a data silo landscape that is not well-connected. An example is that a researcher seeking all the available information on water quality in a certain area would have to query several portals one by one, and even there, they could not find vital data because of different keywords used (e.g., H2O quality, aquatic chemistry, water pollution) or because of the lack of spatial references.

This is a tense situation in the engineering and geoscientific fields that are specialised. Well logs, seismic surveys, and reservoir simulations are important datasets in petroleum engineering worth investing millions of dollars in. When these datasets are disposed of in repositories that store them with bad metadata, then their reuse is crippled, resulting in wasted cost of acquisition and reduced project completion times. On the same note, in mining engineering, information pertaining to mining assays, geotechnical surveys, and resource evaluation is crucial in operational planning and safety. The failure to effectively find and incorporate this information in various projects or firms is a major economic and safety drawback.

This is a key gap that will be filled in this paper through the introduction and validation of the Meta Cheque framework. Meta Check is a fully automated method that can be used to find and fix metadata quality problems on a large scale. As opposed to earlier efforts where the definition of the problem is the main identifying factor, our framework allows users a clear way of remedying the situation and empirically shows the subsequent enhancement in the discoverability of data. We also go beyond generic evaluation to incorporate domain-based profiling and fixing, specifically, high-value areas of petroleum and mining engineering. In this way, we hope to give repository managers and data curators a useful set of tools to make their holdings more useful, thus helping create the truly open, integrated, and powerful global data commons.

Related Work

The metadata quality of open data is not an innovative issue, and a significant research literature has been developed in an attempt to understand its dimensions and offer solutions. Some initial attempts in this direction tended to concentrate on the typology of problems met. In their classical research on European open data portals, Neumaier, Umbrich, and Polleres (2016) have conducted a formal study of the evolution of quality and found that the portals had ongoing problems of runaway missing values in key fields, such as licence and publisher; mixed and uncontrolled vocabularies used in keywords and categories; and a comprehensive failure to meet common metadata standards. Their work listed a standard on how the scope of the problem can be seen in large, heterogeneous data infrastructures.

Likewise, Umbrich et al. (2015) explored the domain of the quality of Linked Open Data (LOD) cloud catalogues, pointing to the problems with the availability and timeliness of metadata as such. They discovered that although metadata is published, it may soon be out-of-date or have broken links, which makes the principle of accessibility an empty idea. All of these studies created an image of an

ecosystem grappling with the fundamentals of data curation, often because metadata development is decentralised and frequently voluntary.

The community has responded to the challenges by coming up with some significant standards and guiding principles. A key standard that was developed by the World Wide Web Consortium (W3C) is known as Data Catalogue Vocabulary (DCAT). DCAT offers a vocabulary in RDF format that is meant to support the interoperability of data catalogues. It provides a standardised format for describing datasets and data services and thus allows automatic aggregation of metadata across multiple sources, which can be used to support federated search portals. But, as we will see, it is not always adopted, and its correct usage involves a degree of expertise which is not always available to publishers of data.

The most significant guidance in the present-day world could be the FAIR Guiding Principles(Wilkinson et al., 2016). FAIR is a brand name used to denote Findable, Accessible, Interoperable, and Reusable. These postulates have been broadly used as the standard of good data stewardship. An example of them is the Findable principle, which explicitly demands rich metadata and a persistent identifier. The principle of interoperability demands the metadata in formal, accessible, shared, and broadly applicable languages and vocabularies. Although FAIR has been offering an effective rallying point and a set of aspirational goals, it is not a technical specification. Its interpretation and application are quite varied and have created a gap in FAIR-ness in which the spirit of the principles is recognised, yet practical implementation is lacking.

In addition to these general frameworks, there have also been studies on individual measurements of metadata quality. Dimensions that are usually assessed are:

- **Completeness:** The percentage of metadata fields that are filled
- **Accuracy:** The rightness of the information that is given in the metadata fields.
- **Consistency:** Lack of logical contradictions in the metadata and usage of standardised format.
- **Timeliness:** How much the metadata is and represents the state of the dataset.
- **Provenance:** The data regarding the provenance of the data, as well as the provenance of the metadata.

Although these metrics are established, the past research has tended to utilise them in a diagnostic and fixed fashion. They respond to the question: What is so bad about it? But rarely go as far as "How can we make it a remedy, and what is the quantifiable advantage of so doing? More so, these domain-specific subtleties of metadata have not received attention. In petroleum engineering, a keyword such as permeability is very specific and critical in meaning, which is lost in a general-purpose thesaurus. Especially, the fields with special metadata needs, such as geoscience (spatial reference systems, geological time scales, and other instrument data), are not commonly supported by generic quality frameworks.

The Meta Cheque framework is developed on this basis. We follow the quality dimensions that have been put in place and align our profiling to FAIR principles. The main contributions of the paper are: (1) a combination of automated

profiling and the collection of realistic, automated repair methods; (2) empirical testing of the effects of these repairs on a important user task- cross-repository search; and (3) the scale up of this approach into the high-value, critical area of engineering geosciences, and indicating thus a way forward to other highly recommended fields.

METHODOLOGY

The Meta Cheque framework was introduced and tested using a systematic four-step protocol that was aimed at being scalable, reproducible, and empirically based. Each phase is described in the subsections below.

1. Harvesting and Creation of Metadata

In order to make the analysis representative, a corpus of metadata representing many open data portals was built. A stratified selection was made to have:

1. Data.gov (United States), European Data Portal (European Union), and Data.gov.uk (United Kingdom) are general-purpose national portals.
2. General-Purpose Research Repositories: Zenodo and Figshare.
3. Domain-Specific Repositories: The EDX (Energy Data Exchange) of petroleum and energy data of the National Energy Technology Laboratory and the USGS ScienceBase Catalogue of the more generalised geoscientific and mining data.

The harvesting was done in a programmatic fashion using Python scripts that connected to the public APIs of such portals. With CKAN-based portals (e.g., Data.gov), we relied on the ckanapi library. In case of repositories that do support OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), like Zenodo, we relied on the use of the library called Sickle. Where bulk download proved to be most effective, we downloaded metadata packages and offline processed them. All the available metadata fields of a dataset were gathered in the course of the harvesting process. Overall, a period of two months, metadata of 52,147datasets was gathered. This metadata was then extracted, and HTML tags were removed where needed and normalised into a single set of JSON schema to aid in later analysis. The schema contained such common core fields as title, description, keywords, tags, licence, publisher, date of publication, temporal coverage, and spatial coverage.

2. Multi-Dimensional Quality Profiling

A strict, automated quality review as a collection of measurable metrics was applied to the normalised metadata corpus. A quality score was computed on several dimensions, in each case of the data set, and executed as a sequence of Python functions.

Completeness (C): This was calculated by the ratio of the number of populated mandatory fields to the number of mandatory fields. Fields that we marked as mandatory were the title, description, keywords/tags, and licence. A field was said to have been populated when it had non-whitespace characters.

C = (Number of Populated Mandatory Fields)/ (Total number of Mandatory fields).

Consistency (Con): This measure was used to assess the syntactic and semantic consistency of field values. It was disaggregated into sub-metrics:

- Date Format Consistency: The proportion of date items (publication, modification) that were in conformance with the ISO 8601 standard.
- Licence Consistency: The proportion of the number of licence fields that could be translated to a standardised SPDX licence identifier, compared to free-text descriptions.
- Geospatial Consistency: The fraction of the spatial coverage fields that might be converted into a standard geometry such as GeoJSON or WKT (Well-Known Text).

This gave an overall consistency score, calculated as an average of these sub-metrics with weights.

Timeliness (T): This indicator measured currency of the metadata. In cases where the datasets had a field of last updated, we determined the number of days passed since the last update. The data sets that were updated within the past year had been considered as timely, the past one to three years old as ageing, and the past more than three years old as stale.

Interoperability & FAIR Compliance (I): This dimension measures the machine-actionability preparedness and system integration.

DCAT Compliance: We verified the presence of the key DCAT elements of the underlying data model, such as dct: title, dct: description, dct: licence, and dcat keywords.

Vocabulary Usage: We quantified the percentage of keywords that were retrieved in a large, aggregated thesaurus of both general-purpose (e.g., EuroVoc) and domain-specific (e.g., NASA GCMD) vocab. Geography: We measured the percentage of keywords retrieved in a large aggregate thesaurus of general-purpose (e.g., EuroVoc) and domain-specific (e.g., NASA GCMD) vocab (Deng, 2013).

Unique Identifier Presence: We determined the proportion of datasets allocated a Persistent Unique Identifier (PID), which was a DOI (Digital Object Identifier) or a specific URI.

The product of this step was a detailed profile of the quality of all the datasets and the corpus as a whole, pointing out the most frequent and critical weaknesses.

3. Auto Repair Solutions

According to the profiling, we have a set of automated repair methods that were applied to the most common problems. They were supposed to be lightweight; that is, these repairs did not need much human effort and could be executed on a large scale.

1. Vocabulary Alignment and Enrichment: This was one of the major repairs to make discoverability better. We used a multi-step procedure for repairing the keywords.

- Tokenisation and Lemmatisation: The NLTK library was used to extract free-text keys and terms in the title and description, tokenise and lemmatise them to their root forms.

- Similarity Matching: The lemmatised terms were then compared to concepts in the controlled vocabularies. As in the case of general repositories, WordNet plus EuroVoc were combined. In the case of the petroleum and mining engineering case study, we utilised NASA GCMD Earth Science Keywords, as well as a custom-built lexicon of terms based on the Petroleum Industry Data Exchange (PIDX)standards and common mining terms (e.g., the EarthChem vocabulary of geochemistry).
- Similarity Calculation: To identify the closest concept in the controlled vocabulary, we combined syntactic similarity (Jaro-Winkler distance) and semantic similarity (based on pre-trained word embeddings such as GloVe).
- Enrichment: The most suitable controlled term was then inserted into the metadata of the dataset as a new and standardised keyword, although the original free-text keyword was retained to provide context.

2. Field Normalisation: It was used to overcome the problem of consistency.

- Date Parsing: We have used the Python dateutil parser to understand an enormous number of date formats and translate them to a common ISO 8601 (YYYY-MM-DD) format.
- Normalisation of licence: A table was built to replace frequent streams found in the free-text licences (e.g., Creative Commons Attribution, CC By) with their canonical SPDX identifiers (e.g., CC-BY-4.0).
- Geospatial Parsing: In the case of spatial fields, we processed spatial coordinates using regular expressions and the shapely library to interpret and verify them and convert them to a standard WKT point/polygon representation.

3. Missing Value Inference: The information provided by the title and description fields is not complete, so, as a way to deal with it, we made simple natural language processing (NLP) inference on the missing values.

- Keyword Inference: To extract the most salient keywords of the description with the help of TF-IDF (Term Frequency-Inverse Document Frequency) and TextRank algorithms, we added them to the keywords field in case it was empty or sparse.
- Category/Topic Inference: Naive Bayes: A naive, simple classifier was used to model a subset of manually categorised data to classify and recommend high-level categories (e.g., Geoscience, Environment, Transportation) based on the text of the title and description.

4. Federated Search Simulation Impact Evaluation.

- The Unity test of metadata quality is how it affects the tasks of the end users. In order to test this, we modelled a federated search environment, which is one of the typical applications of open data (Shen et al., 2024).
- Index Construction: We have constructed two distinct search indices on Elasticsearch version 7.x:
- Baseline Index: Did not include modified metadata, but just the initial metadata of the harvested datasets.

- Repaired Index: The metadata is maintained within the application of all the automated repair techniques.

5. Query Set and Relevance Judgement: A query test query set was created with 25 queries. These included:

- Broad questions: e.g., air quality in New York 2020, schedules of transportation availability.
- Domain Specified Compounds (Petroleum Engineering):e.g., domain: gam ray log Permian Basin, domain: 3D seismic survey North Sea, domain: core sample porosity data.
- Domain-Specific Query (Mining Engineering):e.g., copper assay results in Chile, open pit slope stability, rare element deposit GIS data, etc
- A human curator (ground truth list of relevant datasets) of the entire corpus per query was done automatically in the technical queries. This was not only a labour-intensive move but also a very important step to a proper evaluation.

6. Performance Metrics: We searched the Baseline Index and Repaired Index on the same default Elasticsearch ranking algorithm on each query. The 20 best query results of each query were checked against ground truth. The following standard measures were computed by us:

- Precision: The ratio of the relevant data sets out of the retrieved datasets of the best K results. (We used K=10).
- Recall@K: A ratio of all the relevant datasets in the corpus that were ranked in the top K results.
- F1-Score@K: The harmonic average of Precision and Recall, which gives out one balanced measure.
- Mean Average Precision (MAP): The measure of quality, expressed as a single figure, of all the levels of recall, and is particularly sensitive to the rank of relevant results.
- Mean values of these measures across all queries of the two indices were compared to measure the improvement owing to the metadata repairs.

RESULTS

The application of the Meta Check framework yielded a wealth of quantitative and qualitative results, clearly illustrating both the scale of the metadata problem and the efficacy of our proposed solutions.

Profiling Results: A Landscape of Deficiency

The initial quality profiling painted a stark picture of the state of metadata in open data repositories. The aggregate scores across the entire corpus of 52,147 datasets were low, confirming the hypothesised quality crisis.

Completeness: The average completeness score was 0.58 (on a 0-1 scale). While title (98% populated) and `description` (85% populated) were generally present, critical fields for discovery and reuse were frequently missing. Keywords/tags were missing in 41% of datasets, and license information was absent in a staggering 55% of cases. This means that for a majority of datasets, their terms of

use are unclear, and their topical classification is inadequate. Consistency: The overall consistency score was 0.42. Date formats were highly inconsistent, with only 35% of dates conforming to ISO 8601. License information was particularly problematic; only 20% of the provided license strings matched a standard SPDX identifier, with the rest being free-text descriptions of varying clarity. Geospatial information, when present, was in a persuable standard format only 30% of the time.

Timeliness: Analysis of the `last updated` field (available for 60% of datasets) revealed that 28% of datasets were "stale"(not updated for over three years). This raises serious concerns about the ongoing maintenance and relevance of a significant portion of the open data landscape.

Interoperability: FAIR compliance was low. Only 15% of datasets in general repositories used keywords from a recognised controlled vocabulary. The use of PIDs was higher in research-focused repositories like Zenodo (90%+ have DOIs), but was virtually non-existent in many government portals (less than 5%).

Repair Effectiveness: Significant Quality Gains

The first quality profiling created a bleak image of metadata in open data repositories. The overall scores of the whole set of 52,147 datasets were low, which indicates the assumed quality crisis.

Completeness: The mean score was 0.58 (out of 0-1). Whilst the presence of the critical fields, such as title (98% populated) and description (85% populated) was the rule, the essential fields of discovery and reuse were often absent. In 41% of datasets, there were no keywords/tags, and licence information was missing in an unbelievable 55%of cases. This implies that in most datasets, the terms of use are ambiguous, and their topical classification is unsatisfactory (Wagner & Székely, 2010).

Consistency: The mean score in consistency was 0.42. Dates were extremely incoherent, as only 35 per cent of the dates met ISO 8601. There were also issues with licence information, with only 20 per cent of the given licence texts being identical to a standard SPDX identifier, and the rest being free-text descriptions of varying clarity. The presence of geospatial information was in a standard form, parsable only 30 per cent of the time.

Timeliness: The last updated field (included in 60% of datasets) was analysed to show that 28% of datasets were stale (not updated within the past three years). This casts grave doubts on the continued maintenance and usefulness of much of the open data environment.

Interoperability: FAIR compliance was poor. In general, repositories contain only 15 per cent of data sets that contain keywords in a recognised controlled vocabulary. In research-oriented repositories such as Zenodo (90% of repositories have DOIs), the proportion of PIDs in use was greater, while in most government portals (less than 5% of portals), the proportion of PIDs in use was virtually zero.

Effects on Discoverability: Search-based results

Simulation of federated searches gave the greatest argument in favour of metadata repair. The metrics of performance were all statistically and steadily improving when searching the Repaired Index relative to the Baseline Index.

Metrics	Baseline Index	Repaired Index	Relative Improvement
Precision@10	0.41	0.50	+22.0%
Recall@10	0.32	0.41	+28.1%
F1-Score@10	0.36	0.45	+25.0%
Mean Average Precision (MAP)	0.29	0.38	+31.0%

The enhancement was furthermore greater in the field-specific queries of petroleum and mining engineering. On the query "gamma ray log," recall improved by 0.45 to 0.80 as the non-standard terms (e.g., GR data, natural gamma) started being indexed by the standardised query. In a complex query such as a 3D seismic survey North Sea, the accuracy went up to 0.65, which meant that the user would have very few irrelevant results, and this greatly decreased the time and effort he/she would spend doing the search (Joel et al., 2015). The same increases were found in the mining queries, such as copper assay results, with a substantially better performance of the addition of terms such as Cu analysis and geochemical assay to the formal vocabulary in increasing the recall of the relevant geochemical datasets.

DISCUSSION

The findings of this research are rather conclusive: The open data ecosystem is in a serious crisis of metadata quality that directly hinders its central purpose. Our deep profiling verifies and measures the anecdotal information that has been reported in the previous, smaller-scale research. The extensive lack of license, keywords, and standardised fields cannot be considered a minor inconvenience; it is a fundamental obstacle to data reuse, which results in the existence of a very large landscape of so-called dark data: technically accessible but practically unavailable.

The effectiveness of the Meta Cheque framework proves that such an issue is solvable. The high increase in the quality scores and, above all, the search performance will confirm that automated remediation is a potent and feasible approach. The +25% of F1-score improvement is not a paltry thing; in terms of information retrieval, it is a revolution in user experience. This may be the difference between identifying a critical dataset in minutes and hours, or identifying it and overlooking it altogether for a researcher or engineer.

The case study on petroleum engineering makes a very important point: generic fixes are needed, but not only enough. Geoscientific data is of high value and is technically complex, and thus requires special remediation strategies. Failure to locate a pertinent seismic survey may result in a petroleum company drilling a dry well at an expense of tens to hundreds of millions of dollars. On the same note,

failure to find existing geotechnical information on the stability of a slope may lead to disastrous safety implications in the field of mining engineering (Mao et al., 2024). Our framework starts to meet this requirement with the help of the domain-specific vocabularies such as NASA GCMD and PIDX. The other domain-important intervention is the fix of spatial metadata with the standard coordinate reference systems (e.g., WGS84, UTM) to avoid the expensive mistakes in the spatial analysis.

We do not do it exclusively, however. The methods of repair, though successful, are heuristic in nature. The deduction of the missing keywords is not rooted in deep knowledge, but rather in statistical salience and is sometimes noisy. Novelty There are some terms that may be very new and proprietary and have no equivalents in standard thesauri, which is why vocabulary alignment can become tricky. Also, our present system runs on harvested metadata. To make the repairs permanently effective, they would need to become part of the curation processes of the source repositories, which would demand buy-in on the part of repository managers.

Even though these limitations exist, the way forward is obvious. The value propositions presented through the framework are very compelling to repository managers. It provides a mechanism to increase the visibility and the usefulness of their holdings dramatically with little effort. The principles revealed here can be applied to other high-stakes areas like healthcare, climate science, and materials engineering.

CONCLUSION

This paper has provided an overall evaluation of metadata quality at a large-scale level of open data repositories and proposed the Meta Cheque framework as a viable alternative to systematic profiling and remediation. We have demonstrated beyond a reasonable doubt that the core problems of incomplete metadata are one of the biggest obstacles to the open data vision, and that said issues can be resolved by means of dedicated automation.

The situation can be summarised in 3 lessons:

1. The Problem is Systemic: Metadata quality problems are extensive and dire as well, in fundamental areas needed to support discovery and reuse, and domain-specific repositories have even greater and more complicated problems.
2. Repair in Automated Mode is a Practical and Good Idea: Lightweight methods of vocabulary alignment, field normalisation, and inference of values can generate significant metadata quality changes which directly map to significant improvements in cross-repository search performance.
3. Domain-Specific Context is Essential: To maximise the utility of data in specialised domains, such as petroleum and mining engineering, remediation activities have to use domain-specific standards and vocabularies.

The future work will take into account a number of promising directions. Firstly, we will extend the repair capabilities of the framework to more advanced NLP models that will perform the semantic understanding and value inference more

accurately. Second, we are going to create a domain-adaptable and adjustable form of Meta Cheque that is capable of identifying the subject area of a dataset and picking the best quality measures, and fixing vocabularies without being configured manually. Third will be the development of format-specific repair rules of common file formats in areas of interest, e.g., the automatic extraction and normalisation of header data in SEG-Y (seismic) files and LAS (well log) files to fill metadata fields. Lastly, we will attempt to package this framework as an open-source application that can be deployed by repository managers to continuously observe and optimise their metadata to transition out of a static snapshot of metadata and into a living, ever-better metadata ecosystem.

This study offers a tangible route to the achievement of the potential of the open data movement since once the issue of metadata problems is identified and scalable solutions are adopted, valuable data resources can indeed be discovered, comprehended, and harnessed to advance innovation and address complex world issues.

REFERENCES

Deng, S. (2013, October 3). *Metadata services*. Academia.edu. https://www.academia.edu/4664880/Metadata_services

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers, and myths of open data and open government. Information Systems Management, 29(4), 258–268.

Joel, M., Sminchak, R., Gupta, N., Ladonna, M., & James, F. (2015). *Development of Subsurface Brine Disposal Framework in the Northern Appalachian Basin*. https://www.wvgs.wvnet.edu/www/coop_rpts/reports/Final_Report-Development_Subsurface_Brine_Disposal_Framework_Northern_Appalachian_Basin-10-06-15.pdf

Mao, J., & Ashkan Jahanbani Ghahfarokhi. (2024). A review of intelligent decision-making strategy for geological CO2 storage: Insights from reservoir engineering. *Geoenergy Science and Engineering*, *240*, 212951–212951. https://doi.org/10.1016/j.geoen.2024.212951

NASA. (2023). Global Change Master Directory (GCMD) Keywords. Earth Science Data and Information System (ESDIS) Project. Retrieved from https://gcmd.earthdata.nasa.gov/KeywordViewer/

Neumaier, S., Umbrich, J., & Polleres, A. (2016). Quality assessment and evolution of open data portals. Semantic Web Journal, 7(2), 193-215.

Open Knowledge Foundation. (2023). Open Data Handbook. Retrieved from https://opendatahandbook.org/

PIDX. (2023). Petroleum Industry Data Exchange Standards. Retrieved from https://www.pidx.org/

Shen, J., Zhou, S., & Xiao, F. (2024). Research on Data Quality Governance for Federated Cooperation Scenarios. *Electronics*, *13*(18), 3606. https://doi.org/10.3390/electronics13183606

Umbrich, J., Neumaier, S., & Polleres, A. (2015). Quality assessment of open data portals. In Proceedings of the 9th International Conference on Semantic Systems(pp 1–8). ACM.

W3C. (2023). Data Catalogue Vocabulary (DCAT) - Version 3. World Wide Web Consortium. Retrieved from https://www.w3.org/TR/vocab-dcat-3/

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), 1-9.