Trustworthy Digital Repository Certification: A Longitudinal Study

Devan Ray Donaldson(✉) and Samuel Vodicka Russell

Indiana University, Bloomington, USA
drdonald@indiana.edu

Abstract. Increasingly, government policies are directing federal agencies to make the results of federally funded scientific research publicly available in data repositories. Additionally, academic journal policies are progressively recommending that researchers deposit the data upon which they base their articles in repositories to ensure their long-term preservation and access. Unfortunately, having the necessary technical, legal, financial, and organizational resources for digital preservation is a significant challenge for some repositories. Repositories that become certified as Trustworthy Digital Repositories (TDRs) demonstrate to their stakeholders (e.g., users, funders) that an authoritative third party has evaluated them and verified their trustworthiness. To understand the impact of certification on repositories' infrastructure, processes, and services, we analyzed a sample of publicly available TDR audit reports (n = 175) from the Data Seal of Approval (DSA) and Core Trust Seal (CTS) certification programs. This first longitudinal study of TDR certification over a ten-year period (from 2010 to 2020) found that many repositories either maintain a relatively high standard of trustworthiness in terms of their compliance with guidelines in DSA or CTS standards or improve their trustworthiness by raising their compliance levels with these guidelines each time they get recertified. Although preparing for audit and certification adds to repository staff's dockets of responsibilities, our study suggests that certification can be beneficial. Therefore, we advocate for more specific policies that encourage certification and the use of TDRs.

Keywords: Trustworthy Digital Repositories · Certification · Core Trust Seal

1 Introduction

Increasingly, government policies are directing federal agencies to make the results of federally funded scientific research publicly available in repositories that provide stewardship and access to data without charge while also requiring researchers to better account for and manage these data [11,13,19,20]. Additionally, whether data result from federally funded research or not, academic journal policies are progressively recommending that researchers deposit the data upon

I. Sserwanga et al. (Eds.): iConference 2023, LNCS 13972, pp. 552–562, 2023.
https://doi.org/10.1007/978-3-031-28032-0_42

which they base their articles in repositories to ensure their long-term preservation and access [3,7,25]. Unfortunately, having the necessary technical, legal, financial, and organizational resources for digital preservation is a significant challenge for some repositories [1]. Repositories that become certified as Trustworthy Digital Repositories (TDRs), "demonstrate to both their users and their funders that an independent authority has evaluated them and endorsed their trustworthiness" [5].

To understand the impact of certification on repositories' infrastructure, processes, and services, we analyzed a sample of TDR audit reports from the Data Seal of Approval (DSA) and Core Trust Seal (CTS) TDR certification programs, as they represent the most widely adopted certification programs worldwide, and they make their audit reports publicly available in English. This first longitudinal study of TDR certification over a ten-year period (from 2010 to 2020) found that many repositories either maintain a relatively high standard of trustworthiness in terms of their compliance with guidelines in DSA and CTS standards or improve their trustworthiness by raising their compliance levels with these guidelines each time they get recertified. Although preparing for audit and certification adds to repository staff's dockets of responsibilities, our study suggests that certification can be beneficial. Therefore, we advocate for more specific policies that encourage certification and the use of TDRs.

Although there are currently over 2,400 scientific data repositories covering a broad range of disciplines [22], only a few hundred are certified as TDRs. Some suggest that presently there are not enough policies in place that require certification and use of TDRs to close this gap [16]. While some government policies and academic journal policies require or recommend that researchers make data publicly available [13,19,23], few of these mention TDR standards, certification, and the use of TDRs specifically (c.f., [3]). This is important because data sharing infrastructure networks such as the Common Language Resources and Technology Infrastructure (known as CLARIN), the Consortium of European Social Science Data Archives (CESSDA), and the European Research Infrastructure Consortium for the Arts and Humanities (known as DARIAH) all provide evidence of the power of policy to drive increases in certification as becoming a TDR is a prerequisite for inclusion in and financial support from these networks, and consequently TDR standards such as the Core Trust Seal (CTS) have seen recent increases in applications from repositories, archives, and data centers that wish to join these networks [17].

Besides the benefits of membership in data sharing infrastructure networks and complying with government and academic journal data policies, prior research has explored the benefits that repositories seek via certification. These include: stakeholder confidence, where repositories' funders, the people who deposit data in repositories, and those who use those data will be more confident in repositories' protection and management of the data because they are certified as TDRs; improvements in processes, where conducting self-assessment and audit stimulates repositories to improve their processes and procedures and move to a higher level of professionalism, with an incentive to improve their oper-

ations over time; and transparency, where certification is designed to provide an open statement of repositories' evidence enabling anyone to evaluate repositories' operations and policies [8,9,17]. In contrast, studying the long-term benefits of certification including recertification may prove useful for spurring more repositories to become certified and provoking the development of more policies that require certification and the use of TDRs.

2 Methods

To assess the impact of certification on TDRs, we analyzed 175 audit reports of 127 repositories, 36 of whom got recertified either once or twice. The repositories span five continents and over 26 countries. We selected these repositories because they were certified by the Data Seal of Approval (DSA) and/or its successor, the CTS, the two most widely adopted TDR standards. Both certification programs require a self-audit report that is later reviewed and approved by the standards' representatives. Each audit report describes a repository's level of compliance with a set of 16 guidelines covering a repository's background information, organisational infrastructure (e.g., mission, licenses, continuity of access, sustainability, confidentiality/ethics, skills and guidance), digital object management (e.g., integrity, authenticity, appraisal, storage, preservation, quality, workflows, discovery, identifiers, re-use) and technology (e.g., technical infrastructure and security). We processed all of these documents as a dataset to obtain findings for the measurement of document similarity between recertifications, and to compute term frequency-inverse document frequency (TF-IDF) weights for keyword and topic discovery. Because our focus was on the effects of recertification, we compared the audit reports of all repositories that got recertified, examining the following features: changes in cumulative compliance scores; the number of recertifications; document similarity; and vocabulary terms added and deleted from successive documents.

2.1 Study Design

The purpose of this study was to detect and interpret the significance of changes between chronologically subsequent documents belonging to particular data repositories and their improvement or maintenance of compliance to TDR standards. Natural language processing and topic modeling techniques were employed for two reasons. First, to establish whether changes in documents reflected changes in repositories' overall level of compliance. Second, to extract information, represented as topics (i.e., vectors of tokens), about what changes were being implemented by these repositories.

2.2 Nature of Corpus

The corpus is the entire set of self-assessment audit documents from the DSA and CTS certification programs as of October 2020. All the documents in the

corpus follow the same format of a numerical score and narrative description of a repository's compliance with each of the 16 guidelines. Although both have 16 scored sections, the guidelines for the earlier DSA and more recent CTS certification programs differ in the thematic arrangement of subtopics per section.

The changes in document structure over time led us to pursue methods that would facilitate topic discovery and document similarity comparison on the basis of a "document" being defined as each audit report. However, our acquisition and preprocessing of the dataset allowed us to retain reference to the section-by-section text and numerical scores of each document to facilitate the discovery of clusters of topics that demonstrate different rates of change and stability across the 2–4–year intervals between recertifications.

The size of the corpus was relatively small (n = 175) though each document contained at least 1,000 words.

2.3 Data Acquisition

Audit reports were obtained from two sources. First, we acquired all DSA and early CTS audit reports from a MySQL database archived and made accessible in DANS EASY [6]. Second, we acquired more recent CTS audit reports from the list of certified repositories on the CTS website [4].

We migrated and extracted the audit reports and their metadata from both sources into a file-based SQLite database that would serve as inputs for analysis. Our database [10] includes the section-by-section text and numerical scores of each repository's audit reports, along with information used to identify repositories.

To arrive at our sample, we filtered raw data based on three criteria. First, to only include audit reports that had both a numerical score and a response text entry for each of the guidelines. Second, we de-duplicated the audit reports so that each repository had either zero or one audit report for each certification period. Third, to identify the subset of repositories that recertified either once or twice between 2010 and 2020, we ran queries on our database.

2.4 Models and Data Analysis Techniques (Feature Selection)

To process and analyze the data derived from raw text, we used multiple techniques: rule-based systems for text-preprocessing; a pre-trained vector space model for word embedding to compute document similarity comparisons; term-frequency inverse document-frequency (TF-IDF) to refine token collections; and latent Dirichlet allocation (LDA) to produce a topic model.

We used the Python NLP library SpaCy [14] to provide a suitable word vector model and utilities for preprocessing. We used the large English language model package [24] obtained from SpaCy's pre-trained model download script. This model package implements methods for part-of-speech parsing, named entity recognition, and lemmatization based on a convolutional neural net trained on

OntoNotes 5.0 dataset. Also included in this package is the Common Crawl-trained GloVe word vector model which we used to analyze document content quantitatively.

The baselines of average improvement and/or maintenance of compliance for comparison against the results of our topic analysis were established by obtaining descriptive statistics for the sum of the compliance level scores (ranging from 0 to 4) reported for each section within each TDR audit report. We also found the slope of the least-squares linear regression for these cumulative compliance level scores for repositories that recertified at least once.

To quantitatively compare document text and to prepare the dataset for topic modeling, we used Python scripts to read the document string data from the SQLite database into the SpaCy language processing pipeline. To quantify the degree of differences between documents, we computed similarity scores, which represent cosine similarity, obtained by finding the Euclidean distance of the L2 vector norm applied to the dot product of each document's tokens. We also created lists of uniquely added and removed terms for all cases of recertification by finding the set difference of the lemmatized form of sets composed of each token from the earlier and later documents. These lists were combined with contextual information identifying the repository, the report, the token's vector norm, document similarity, cumulative score, etc. to aggregate the relevant tabular data in a single flat file.

After constructing our comprehensive table of document changes, we created histograms to visualize the extent to which changes in content reflect changes in TDRs' cumulative compliance level scores.

2.5 Topic Discovery Techniques

In addition to cumulative score and document similarity, we examined whether these changes coincided with topics discussed in the documents. We used part-of-speech, regular expressions, and other rule-based utilities provided by Python and SpaCy to filter out "noisy" tokens.

We also used the Python libraries Matplotlib [15] and SciKit-Learn [21] to visualize word distances of terms frequently added or removed from the documents. We used the Principal Component Analysis (PCA) algorithm supplied by SciKit-Learn to decompose the representative 300-element word vector of each term into a 2-dimensional point, along with the k-means clustering algorithm provided by SciKit-Learn to examine how the terms group together. To select input values for PCA, we sorted the list of words by their TF-IDF weight into three categories: highly specific terms (high-weight); an intermediate group; and broadly general terms (low-weight). For these TF-IDF categories, we selected the 20 most frequently occurring terms. We then used the LDA model from SciKit-Learn to generate a representation of changes in document content derived directly from our corpus.

The parameters required for LDA include: number of topics; number of passes and iterations to be performed; and the alpha and beta parameters for expected topic-document density and topic-word density [2,12]. Because we did not have

any prior expectations about the topic-document density and topic-word density, we used the default arguments of 1/number of topics (n = 8) for the priors. Sentences associated with terms that changed were loaded into a sparse matrix and transformed by the LDA model into a distribution of topics represented in the sentence. We selected the top three proportionally most representative topics for each sentence. For both groups–terms classified by PCA and k-means, and terms classified by LDA–we found the mean rate of change in cumulative compliance score by referencing the rows in our document changes table that contained those tokens. We also used the terms changed data as an aid for finding examples of improvements as demonstrated by text added and text deleted for a repository whose cumulative compliance level score significantly improved after recertification.

3 Findings

3.1 Cumulative Compliance Scores

Analysis of descriptive statistics for the TDRs' compliance level scores shows that repositories that recertify commonly report both increases, and to a lesser extent decreases in their compliance, with the mode amount of change being +2.5. Performing a least-squares linear regression on the scores of repositories that recertify shows a slope of 0.08, bearing a slightly positive trend (see Fig. 1). We observed a ceiling effect where most of the TDRs' cumulative compliance scores cluster near 64, the top of the graph and the maximum cumulative compliance level for these TDRs (see Fig. 2). Additionally, analysis of the data along the x-axis demonstrates that most of the repositories' scores change minimally, that is, no more than a gain or loss of five points between certifications.

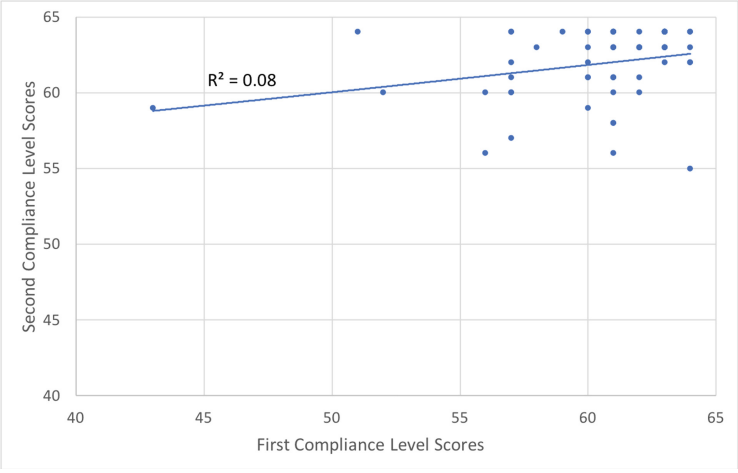

Fig. 1. Changes in TDRs' compliance level scores.

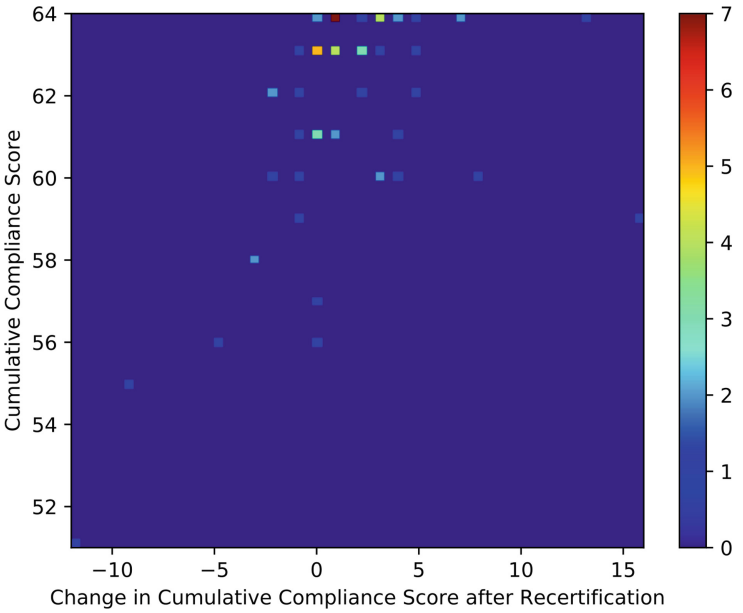

Fig. 2. Changes in compliance with TDR standards. This heatmap shows changes in repositories' compliance with TDR standards each time they recertify. The colors reflect how many repositories had similar compliance level score changes (n = 36).

3.2 Document Similarity Comparisons

As shown in Fig. 3, we found a correlation between the document similarity comparisons obtained with word vector modeling and the amount of change observed between reported compliance scores from repositories' subsequent recertifications. Taken together, these findings suggest that when TDRs' numerical scores change, the text in their audit reports also change to a similar degree. We found the set difference of vocabulary terms per document to contain the addition of 36,328 words and the removal of 8,675 words.

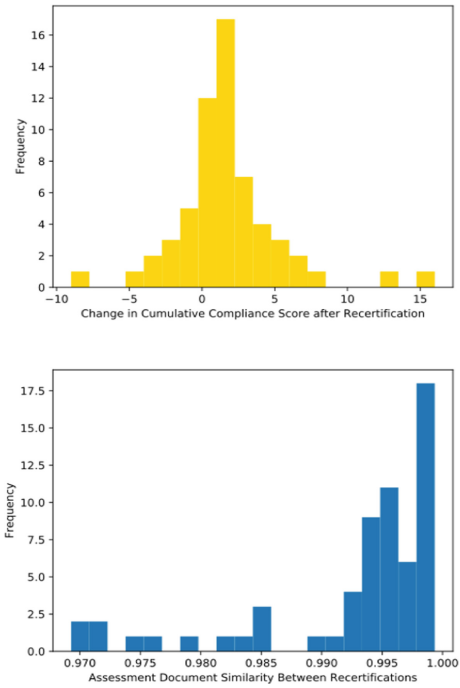

Fig. 3. Histograms comparing TDRs' audit reports. These comparisons consider consecutive recertification (e.g., comparing 2010 certification to 2014 recertification or 2014 certification to 2017 recertification) and non-consecutive recertification (e.g., comparing 2010 certification to 2017 recertification) for repositories that got recertified twice between 2010 and 2020 (n = 36). The top histogram compares cumulative compliance level scores of repositories showing that repositories' change in score based on recertification typically ranges from −2 to +4, with a tail extending to both extremes (from −9 to +16). The bottom histogram compares document similarity of audit reports from consecutive and non-consecutive recertifications showing a concentration around a small degree of difference with a tail extending towards 0, which contains both negative and positive extremes of the difference in scores between certifications (from 0.96 to 0.99).

3.3 Topic Modeling

As shown in Fig. 4, the results of transforming passages of changed text with a topic model show that most of the changes to audit reports when repositories recertified correspond to five of our topics: governance, organizational networks and expertise (Topic 3); fitness-for-use of data by researcher communities (Topic 2); security and recovery planning (Topic 6); licensing and ethics (Topic 4); and discovery and reuse of data by end-users (Topic 1). The topics less likely to be the subject of textual changes were associated with our remaining three topics: versioning, integrity, description, and metadata harvesting (Topic 0), requirements,

standards, and best practices for metadata, file formats, deposit, and submission (Topic 5); and infrastructure, workflows, and interfaces for data lifecycle management (Topic 7).

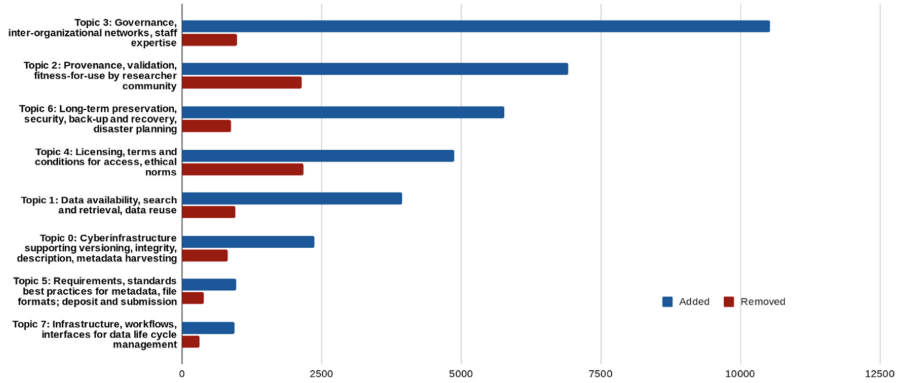

Fig. 4. Topic frequency in change text.

3.4 Improvements

For repositories whose cumulative compliance scores changed the most between certifications (i.e., scores improved by 10 or more points), we identified improvements to their storage, quality control processes, codes of conduct, workflows, cyberinfrastructure, and their adoption of other relevant repository standards. For example, one repository reported no evidence of compliance in multiple areas the first time they certified, and in contrast, reported full compliance for those guidelines when they recertified.

The finding that text associated with depositor requirements was poorly represented among changes in document vocabulary may indicate greater sophistication of both computational and human systems for accessioning data of increased variety in quality and formats for TDRs over time. Although, at the surface level, it might seem counterintuitive to associate accessioning data, including those that range in quality, with improvement, in reality, if a repository can preserve data of less-than-perfect quality, it is better than the data not being preserved at all. Furthermore, preserving data of varying levels of quality requires a metadata strategy capable of reliable data quality representation. Standards and requirements for deposit continue to be important for digital preservation, but an increased focus on data description and quality assessment implies an improvement for different classes of stakeholders, for example, with more flexibility for data producers and greater assurance for data consumers.

In sum, our analysis of ten years of repositories' DSA and CTS audit reports suggests that these TDRs are discussing exactly the types of topics that are vital

for data management and sharing. Our findings demonstrate that these repositories expanded their purview in response to digital preservation challenges beyond bit-level fixity with strategies for long-term organizational sustainability to focus on maximizing their holdings' accessibility and usefulness for researchers. Moreover, our results show that many of these TDRs have either maintained a standard of excellence or have improved in their stewardship capabilities as a result of recertification. Topic frequency in changed text was more distinct among words added than words removed, suggesting that improvement is expressed by developing new services and strategies for continued access and preservation, while less drastic revisions are evidence of maintenance of existing capacity.

4 Recommendations

We found that repositories in our sample either maintained or increased their compliance with DSA or CTS TDR standards over time. Since attaining certification involves third-party evaluation of a repository's capacity and commitment to preserving and making data publicly available [18], we offer the following recommendations based on our results. First, we recommend that policymakers who mandate open access to the results of federally funded scientific research revise and expand their directives to include explicit verbiage about certification and the use of TDRs. Specifically, funders should require data repositories to undergo audit and attain certification by CTS or some other certifying body. And funders should require or recommend that their grantees deposit data in TDRs. Second, we recommend that more journal policymakers update their data policies to require authors to deposit their data in TDRs. Even though we are starting to see these trends [3,11], more policy needs to be developed in this area.

References

1. Austin, C.C., Brown, S., Fong, N., Humphrey, C., Leahey, A., Webster, P.: Research data repositories: review of current features, gap analysis, and recommendations for minimum requirements. IQ **39**, 24–38 (2016)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. JMLR **3**, 993–1022 (2003). https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
3. Coalition for Publishing Data in the Earth and Space Sciences (COPDESS) - Author Guidelines. http://www.copdess.org/enabling-fair-data-project/author-guidelines. Accessed 16 Sept 2022
4. CoreTrustSeal, Core Certified Repositories. https://www.coretrustseal.org/why-certification/certified-repositories/. Accessed 12 Aug 2022
5. CoreTrustSeal Standards And Certification Board: CoreTrustSeal Trustworthy Data Repositories Requirements 2020–2022 (2019). https://doi.org/10.5281/ZENODO.3638211
6. Data Seal of Approval (DSA) - EASY. https://easy.dans.knaw.nl/ui/datasets/id/easy-dataset:116038/tab/2. Accessed 12 Aug 2022

7. Data Policies|Scientific Data. https://www.nature.com/sdata/policies/data-policies. Accessed 16 Sept 2022
8. Dillo, I., De Leeuw, L.: Ten years back, five years forward: the data seal of approval. IJDC **10**, 230–239 (2015). https://doi.org/10.2218/ijdc.v10i1.363
9. Donaldson, D.R., Dillo, I., Downs, R., Ramdeen, S.: The perceived value of acquiring data seals of approval. IJDC **12**, 130–151 (2017). https://doi.org/10.2218/ijdc.v12i1.481
10. Donaldson, D.R., Russell, S.V.: Replication Data for: "Trustworthy Digital Repository Certification: A Longitudinal Study", Harvard Dataverse (2022). https://doi.org/10.7910/DVN/TDX2J8
11. European Commission, Directorate-General for Research and Innovation: Guidelines on FAIR data management in Horizon 2020 (2016)
12. Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference. JMLR **14**, 1303–1347 (2013). https://www.jmlr.org/papers/volume14/hoffman13a/hoffman13a.pdf
13. Holdren, J.P.: Increasing access to the results of federally funded scientific research. Executive Office of the President, Office of Science and Technology Policy, Washington, D.C. (2013)
14. Honnibal, M., Montani, I.: spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017)
15. Hunter, J.D.: Matplotlib: a 2D graphics environment. Comput. Sci. Eng. **9**, 90–95 (2007). https://doi.org/10.1109/MCSE.2007.55
16. Husen, S., de Wilde, Z., de Waard, A., Cousijn, H.: Recommended versus Certified Repositories: Mind the Gap, https://papers.ssrn.com/abstract=3020994, (2017). DOI: https://doi.org/10.2139/ssrn.3020994
17. L'Hours, H., Kleemola, M., De Leeuw, L.: CoreTrustSeal: from academic collaboration to sustainable services. IQ **43**, 1–17 (2019). https://doi.org/10.29173/iq936
18. Lin, D., et al.: The TRUST principles for digital repositories. Sci Data. **7**, 144 (2020). https://doi.org/10.1038/s41597-020-0486-7
19. Marcum, C.S., Donohue, R.: Breakthroughs for all: delivering equitable access to America's research. https://www.whitehouse.gov/ostp/news-updates/2022/08/25/breakthroughs-for-alldelivering-equitable-access-to-americas-research. Accessed 16 Sept 2022
20. Obama, B.: Executive Order - Making Open and Machine Readable the New Default for Government Information. https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-. Accessed 16 Sept 2022
21. Pedregosa, F., et al.: Scikit-learn: machine learning in python. JMLR **12**, 2825–2830 (2011)
22. Re3data.org. https://www.re3data.org. Accessed 19 Sept 2022
23. Research Data - Elsevier. https://www.elsevier.com/about/policies/research-data. Accessed 16 Sept 2022
24. spaCy Models Documentation. https://spacy.io/models/en. Accessed 12 Aug 2022
25. Understanding and using data repositories. https://authorservices.taylorandfrancis.com/data-sharing/share-your-data/repositories. Accessed 16 Sept 2022