Data Stewardship Exercise 2

This project was created for the course Data Stewardship UE on the TU Wien in SS2023.

Introduction

This report analyzes the Core Trust Seal certification task for the Research Data Repository of TU Wien (TUWRD). The focus is on the "Preservation plan (R09)" requirement. The goal is to examine existing self-assessment reports from similar repositories, compare their responses, and provide a complete response for TUWRD, aiming for the highest level of compliance.

Prerequisite

To conduct the analysis, the following inputs were provided:

- Core Trust Seal criteria and extended guidance for 2023-2025.
- Existing self-assessment reports from 10 repositories (similar to TUWRD) based on previous Core Trust Seal guidelines.
- Relevant documents of TUWRD, including Terms of Use, Policies, and FAQs.

Selection Criteria

The selection of examples for analysis was done systematically, taking into consideration repositories that are similar to TUWRD. In this analysis, the reports were selected from the Core Trust Seal page, under certified repositories (https://amt.coretrustseal.org/certificates). The main selection criteria was for being similar to large repositories, typically of universities and large organizations.

To summarize, the following criteria were used for the selection:

- The aim was to gather a diverse range of self-assessment reports that align closely with TUWRD in terms of their characteristics and requirements.
- Repositories with similar types of data (research data from an academic institution).
- Repositories that have undergone the Core Trust Seal certification process previously.
- Repositories that adhere to similar preservation standards and best practices.
- Repositories with a comparable scale of operations and resources.

We tried to select more general repositories, and not those of specific subjects, but this was challenging considering the limited number of certified repositories, therefore, when we decided to choose more subject specific repositories we tried to select larger repositories, and typically research repositories.

Requirements

Please note we were unsure whether we had to analyze all requirements, or only R9, so we chose to analyze all requirements to be certain. However, we focused specifically on requirement R9, which is why it is specifically analyzed for each report. Analyzing each requirement for each report could be possible, but we felt doing this in detail for each criteria was unnecessary considering we are not required to do that for TUWRD, only for the preservation plan.

Core Trust Seal Requirements:

- R0. Background Information & Context
- Organizational Infrastructure
    - Mission & Scope (R01)
    - Rights Management (R02)
    - Continuity of Service (R03)
    - Legal & Ethical (R04)
    - Governance & Resources (R05)
    - Expertise & Guidance (R06)

- Digital Object Management
    - Provenance and authenticity (R07)
    - Deposit & Appraisal (R08)
    - **Preservation plan (R09)**
    - Quality Assurance (R10)
    - Workflows (R11)
    - Discovery and Identification (R12)
    - Reuse (R13)

- Information Technology & Security
    - Storage & Integrity (R14)
    - Technical Infrastructure (R15)
    - Security (R16)

Analysis of Existing Self Assessment Reports

The analysis of the existing self-assessment reports involved examining the responses provided by the selected repositories. The following aspects were considered:

- How is this criterion addressed? The analysis focused on understanding how each repository addressed the preservation plan requirement. It involved examining the repository's approach to data preservation, including strategies, processes, and infrastructure.

- Difference between levels of compliance: The analysis identified variations in compliance levels among the repositories. This included differences in the comprehensiveness of preservation plans, the extent of adherence to best practices, and the level of integration of preservation activities within the repository's operations.

- Evidence provided to support statements: The analysis evaluated the evidence provided by the repositories to support their statements related to the preservation plan. This included documentation, policies, guidelines, preservation frameworks, and any other relevant materials that demonstrate the repository's commitment and implementation of preservation strategies.

- Positive/negative examples: The analysis identified positive examples (when applicable) where repositories showcased exemplary preservation practices, as well as any negative examples where repositories had deficiencies or gaps in their preservation plans. This provided insights into best practices and areas that need improvement.

The analysis of the existing self-assessment reports helped gain a broader understanding of how similar repositories approach the preservation plan requirement and provided valuable insights into successful strategies and potential pitfalls.

Report 1: Edinburgh Datashare

They note specifically for all different requirements what is being done to accomplish the requirements and how it is being done. They often cite their policies and practices by linking relevant information. They provide examples, and use detailed responses to make everything clear. Typically, the differences between compliance levels of 3 and 4 is not drastic, often that there is no formal agreement or the plan laid out is not completely finished at the time of reviewing the repository. For levels 1 & 2 however the differences are more pronounced. However, this repository had no level 1 and 2 scores. The only level 3 score is for requirement R3, and this is only because of their lack of a concrete formal agreement with another repository that the other repository will take over all duties. Typically, each criteria is addressed quite thoroughly using their policies and the supporting documentation to provide additional evidence, along with the written explanation of their level and how they are meeting each criterion. This is particularly detailed in comparison to some of the other reports, but this is particularly helpful in understanding, especially because the reviews commented about how this application is particularly good and thorough. Therefore, I chose to review this one first, as it allows for a good starting point for high levels of detail, clear policies and workflows, and detailed evidence and quotations from other parts of the university and outside organizations. It is also a university organization, similar the the TU Wien, and therefore it allows us to compare similar repository types.

Preservation Plan: They note their system in OAIS terms, and use detailed documentation to allow for clear procedures in regards to data in the repository being analyzed by curators with the DSpace workflow, as well as high-quality user-facing web pages, reviewed annually. Users can either do batch depositing through DSpace, or the more popular web interface, both of which produce high quality SIPs, which are reviewed by curators, and metadata is stored along with data to produce AIPs. Finally, all AIPs will be also deposited into the DataVault for storage so that long-term storage is ensured. Additionally, the metadata is harvestable and is harvested by OAIS-PMH, Google Scholar, Google Datasets, etc. All of this, along with backups and additional measures such as backing up the database to tape daily, database snapshots, a versioned database and more, make sure the data is fully safe, and that the requirements are fully met.

Report 2: Apollo, Cambridge University Library

Again, this was chosen because of the similarity, as it is also a university storage system, run in collaboration with the library which is similar to the TUWRD. It is an institutional repository, as well as a publication repository (unlike Edinburgh DataShare, which is not a publication repository). They do only basic data curation, the same as the Edinburgh DataShare. They have detailed responses to all criteria, as they cite for each section supporting documentation, and reference it throughout their response to the specific criterion. This repository had two level 3 ratings, for R3 and R10. This means it is in the process of implementation, and while there were a few requirements for which they had to be re-reviewed as they clearly resubmitted, they achieved the level 4 rating with clearer text and citations for each section. Their level of detail is again quite high, similar to the University of Edinburgh, and in several cases even higher. I particularly liked the citations of relevant policies and all relevant links being listed for the reviewer to see, which clearly backs up their textual evidence detailing the meeting of all criteria, or the reasons why it is not yet met. This is particularly interesting because the University of Cambridge is typically regarded extremely highly in the academic field. Their repository is similarly of extremely high quality, and their documentation is as well. They note for each criterion how exactly they manage to pass, citing evidence such as “The Apollo repository is built on the latest 5.x version (5.10 at the time of writing) of the DSpace open-source repository platform.”, they then proceed to link to it on GitHub, discuss the DS Ops team, and additional relevant information in excruciating detail, such as how the VM is run, future plans for Apollo software, and more. This level of detail allows the reviewers to clearly see all aspects and quickly check the provided links to scrutinize each requirement.

Preservation Plan: They note the DS Ops team is responsible for the infrastructure underpinning Apollo. The information is available in the team's internal Wiki, and of course is updated and available to other relevant teams and stakeholders. As Apollo is designed for long-term storage, it is well-equipped with monitoring integrity and consistency, and this includes checksumming, the monitoring of the disks using a NetApp application for health, NetApp ONTAP for the deterioration of storage-level media. NetApp SnapMirror makes copies of each piece of data, replicated to a second appliance. There is also a backup and tape management strategy to assist in making sure data cannot be lost, as there are numerous backups and they are also consistently backed up to tape so as not to lose information. They receive alerts from the NetApp ONTAP application when errors are found and the severity of them, and a team is assembled if necessary to handle the incidents. Overall, the storage and management is robust and well thought out, as well as having data recovery policies laid out as well. Thus, the repository fully meets the requirements.

Report 3: CU Scholar, University of Colorado Boulder

This is once again an institutional repository, as well as a library. This was again chosen for its similarity in function to TUWRD, although none of the CTS repositories are a perfect match, this is relatively close as well. They do Data-level curation, which is significantly more detailed than either the DataShare or the Apollo repository do. Reviewing all datasets at this level requires lots of resources and detailed documentation of what is changed. R3, R9, and R10 are all in the process of being implemented for CU Scholar, which is a departure from the previous two, where R3 and R10 were in progress for Apollo, while DataShare had only R3 as in progress. This is especially interesting because R9 is the specific requirement we are focusing on for TUWRD. In addition, we note that the details of why they are only in the implementation phase are of particular interest for comparison. They note the details of each criteria, and then how they are fulfilled, however it is relatively significantly less detailed in terms of citations, in that while the actual text length is similar for each requirement, the number of citations as well as the detail level of the citations in providing other links and general resources is lower. The links are still provided, however, they are included in the citation instead of listing all helpful links underneath citations, as a courtesy to reviewers. This functions quite similarly but is slightly less readable and a bit more confusing when following links, as otherwise links are simply not mentioned in the text. This contrasts with the review of DataShare, where numerous links were included. They give positive examples of how they store data with R10, quote their own terms of use and preservation policy, and generally provide details similar to the level of both Apollo and DataShare, but it is also clear that they have less intention to store data for an undetermined amount of time, as they specifically note that every 10 years data is evaluated for perceived value to researchers and in the case it is no longer considered relevant, will be removed or outsourced from the repository. It also notes that it is in some cases a best-effort attempt, especially for proprietary formats, and that no guarantees can be made for readability. This is interesting as it is similar to what is considered for both Apollo and DataShare, but none are exactly the same.

Preservation Plan: In this case, we see similar cases made with both Apollo and DataShare, as it is stored in AWS ElasticBlocks, S3 and in the PetaLibrary. They also use fixity checksums, similar to the others, to maintain integrity for the data. They are also a member of APTrust, which means they get additional preservation support such as distributed data copies, fixity ingest checks, and quarterly thereafter. However, this is less complete than for example either Apollo or DataShare. There is also no mention of metadata. As such, we find this to be slightly less convincing than either Apollo or DataShare, and we suppose this is the reason for the level 3 rating, as it is not yet at the level expected for full completion of the requirements. The citations are enough, but the lack of checking on media, or the lack of alerts. Generally, the details of this requirement are not nearly as thorough as the previous, and this is interesting to note going forward, when evaluating this requirement. However, it is clearly still preserving data to a high standard.

Report 4: 4TU.ResearchData

4TU.ResearchData is a national facility for storing and preserving science and engineering research data. It was established in 2008 as an initiative of three technical universities in the Netherlands.. The repository has been fully operational since 2010 and has evolved to become a trusted and certified repository. It maintains over 6,800 datasets, corresponding to about 42 TB of data. 4TU.ResearchData actively participates in Research Data Netherlands, a national coalition of data archives, to promote long-term archiving and data reuse. The self-assessment provides evidence to support the statements regarding preservation planning in 4TU.ResearchData. It mentions the repository's policies, guidelines, and procedures for data quality review, metadata completeness, sustainability of file formats, and the presence of readme files. The self-assessment also highlights the adherence to the Findable, Accessible, Interoperable, and Reusable (FAIR) data principles. The evidence includes links to websites, publications, and deposit agreements that outline the repository's practices. Positive examples include the comprehensive data quality checks, value-added curation activities, and the adoption of suitable licenses (including Creative Commons licenses) for dataset sharing. These examples showcase 4TU.ResearchData's efforts to ensure data integrity, accessibility, and adherence to ethical and legal requirements. No negative examples are explicitly provided in the self-assessment. In response to the preservation plan criterion, 4TU.ResearchData provides a comprehensive strategy for long-term data preservation. The repository ensures the sustainability and accessibility of deposited datasets through a deposit agreement that grants permission to store, copy, and modify datasets while preserving their content. It emphasizes the importance of funding for long-term sustainability and highlights the commitment of partner institutions to ensure the availability of data even in challenging situations. The self-assessment also outlines the curation levels and quality assurance procedures followed by 4TU.ResearchData. The repository performs data quality checks, adds metadata and documentation, and works with depositors to address any formatting or content issues. By adhering to these practices, 4TU.ResearchData enhances the quality and usability of datasets, making them more valuable for researchers. Overall, the self-assessment for 4TU.ResearchData demonstrates a robust preservation plan. The repository's commitment to long-term access, data quality checks, curation activities, and adherence to licensing and ethical considerations

strengthens its position as a trusted repository for science and engineering research data. The provided evidence supports the repository's statements and showcases its dedication to preserving and enhancing the value of research datasets.

Preservation Plan: The preservation plan for 4TU.ResearchData focuses on ensuring the long-term accessibility and integrity of the deposited research data. Here is a summary of the preservation plan:

1) Data Integrity: 4TU.ResearchData performs data quality checks during the curation process to ensure the completeness and accuracy of metadata, sustainability of file formats, and the presence of essential documentation. These checks help maintain the integrity of the original data and ensure compliance with repository standards.
2) Preservation Strategy: The repository follows best practices and adopts the Findable, Accessible, Interoperable, and Reusable (FAIR) data principles to improve the sustainability and reusability of datasets. It promotes the use of preferred file formats and provides guidance on responsible data management to researchers.
3) Sustainability and Funding: 4TU.ResearchData relies on adequate and reliable funding from partner institutions to support the long-term preservation of data. The repository has a consortium agreement in place, ensuring the commitment of partner organizations to sustain the repository and preserve the deposited datasets.
4) Disaster Recovery: The repository has established disaster recovery plans that align with the University's data center procedures and policies. These plans undergo regular review and include measures to mitigate the impact of potential disasters, ensuring the continuity of access to the preserved data.
5) Minimum Preservation Period: 4TU.ResearchData guarantees a minimum preservation period of 15 years for published datasets. This ensures that the deposited data will remain accessible and usable for an extended period, supporting ongoing and future research.
6) Data Stewardship: The repository follows data stewardship practices to manage and curate the deposited datasets effectively. Curators review and enhance metadata, provide documentation, and perform necessary transformations to address formatting and content issues, improving the usability and discoverability of the data.
7) Data Migration: 4TU.ResearchData stays abreast of technological advancements and best practices in digital preservation. It commits to migrating data to new formats, platforms, and storage media when required, ensuring that the datasets remain accessible and compatible with evolving technologies.

Overall, the preservation plan of 4TU.ResearchData emphasizes data integrity, sustainability, disaster recovery, and ongoing data stewardship. By following established preservation strategies and engaging in continuous improvement, the repository aims to ensure the long-term accessibility and usability of the research data it hosts.

Report 5: Data Repository for the University of Minnesota

The Data Repository for the University of Minnesota (DRUM) demonstrates its commitment to preservation through various measures. It operates as a well-curated subset of the larger University of Minnesota Digital Conservancy, providing a dedicated space for sharing, publishing, and preserving digital data. DRUM accepts deposits only from University of Minnesota affiliates but ensures that all data housed in DRUM are publicly available. The repository is listed in the Registry of Data Repositories re3Data.org, which enhances its visibility and credibility. The self-assessment does not explicitly mention different levels of compliance for the preservation plan criterion. However, it emphasizes the comprehensive approach taken by DRUM to ensure long-term preservation of data. This includes the signing of a deposit agreement by data producers, granting long-term stewardship rights to the University of Minnesota's Board of Regents. The repository operates under the umbrella of the University Libraries, which assume responsibility for preserving institutional assets, including data in DRUM. This self-assessment provides evidence to support the statements regarding preservation planning in DRUM. It mentions specific policies, agreements, and guidelines that outline the repository's approach to preservation. The Terms of Use policy, End User Access Policy, and University-wide Acceptable Use of Information Technology Resources Policy establish the access and use conditions for data in DRUM. The University Libraries Digital Preservation Framework and disaster recovery plans demonstrate the repository's commitment to continuity of access. Additionally, the curation procedures and collaboration with experts ensure the management of disclosure risk and adherence to ethics and disciplinary norms. Positive examples include the clear communication of policy requirements to depositors, curatorial review of datasets, and the establishment of collaborations with various University committees and research offices. These examples highlight DRUM's proactive efforts to ensure data integrity, authenticity, and ethical use. No negative examples are explicitly provided in the self-assessment.

Preservation Plan: In response to the preservation plan criterion, DRUM outlines its strategy for long-term data preservation in a separate document. The preservation plan includes details on the repository's infrastructure, staff expertise, and data curation workflows. It aligns with best practices in the field of digital preservation, addressing aspects such as data formats, disaster recovery, and compliance with ethical and legal requirements. The plan provides evidence through documentation, policies, and procedures, demonstrating DRUM's commitment to preserving data for long-term access and future use. Overall, the self-assessment for the Data Repository for the University of Minnesota (DRUM) demonstrates a comprehensive approach to preservation planning. The repository ensures long-term accessibility, adheres to relevant licenses and policies, maintains data integrity, and collaborates with experts and University committees to ensure the highest level of compliance. The provided evidence supports the repository's statements and showcases its dedication to preserving valuable research data.

Report 6: The Linguistic Data Consortium (LDC)

As most of the repositories reviewed in this report, The Linguistic Data Consortium (LDC) repository, is also hosted by an academic institution, such as the University of Pennsylvania, and addresses various criteria to ensure effective data management and preservation. Preservation is a fundamental objective, reflected in LDC's practices of identifying, archiving, and storing data to maintain its original form. It creates and distributes language resources, supports sponsored research programs, and collaborates with international organizations. Preservation is a key objective, and LDC's practices ensure that data is identified, archived, and stored to preserve its original form. The repository also focuses on licensing agreements to regulate data use, with different levels of compliance depending on the type of agreement signed by members, licensees, and task participants. Licensing agreements regulate data use, and compliance is enforced through agreements and legal oversight. Evidence supporting these statements includes membership agreements, user agreements, and corpus-specific license agreements. LDC provides positive examples such as its distribution of over 200,000 copies of databases to organizations in over 100 countries. Negative examples might include instances where data cannot be published due to non-compliance with legal or ethical regulations. The repository's continuity of access is supported by the University of Pennsylvania's Van Pelt Library, which catalogs and preserves LDC publications. Evidence includes catalog records and plans for continued accessibility even if LDC ceases operations. Confidentiality and ethics are addressed through a submissions review process, compliance with legal regulations, and the removal of personal identifying information. Evidence includes agreements signed by data providers, evidence of compliance with regulations, and the absence of GDPR issues. The repository operates within the organizational infrastructure of the University of Pennsylvania, supported by funding and staff resources.

Preservation Plan: The preservation plan for the Linguistic Data Consortium (LDC) is comprehensive and ensures the long-term accessibility of all data in the LDC Catalog. The plan covers several key aspects:

1) Long-term Accessibility: LDC is committed to preserving and providing access to all deposited corpora, including those created in the 1980s. There are no differences in preservation levels for published corpora, ensuring equal preservation efforts for all data.
2) Preservation Measures: LDC employs redundant backups, multiple drives, and off-site storage of physical copies of all corpora to ensure long-term preservation. This approach safeguards data against loss and provides resilience against hardware failures or disasters.
3) Migration and Format Obsolescence: LDC follows best practices in the digital preservation community by migrating data to new formats, platforms, and storage media as required. This ensures that data remains accessible even as technologies and formats evolve. Previous versions of data are retrievable, allowing for traceability and comparison over time.
4) Integrity and Fixity Checking: The storage infrastructure used by LDC incorporates ongoing integrity checking mechanisms to ensure the fixity and integrity of the stored data. This helps to identify and prevent data corruption or alteration.
5) Metadata and Distribution Agreements: To facilitate preservation and access, providers are required to submit appropriate metadata alongside their resources. Additionally, all providers must sign a distribution agreement with LDC, granting the Consortium the rights to store, transform, and distribute the submitted resources. These agreements enable LDC to fulfill its critical function of ensuring long-term access and understandability of the data.

The evidence supporting these statements includes URLs to LDC's IT infrastructure, curation and distribution services, publication process, and preserving data. Positive examples are highlighted, such as the availability of corpora deposited since the 1980s and the implementation of redundant backups and off-site storage. The provided URLs offer additional information and documentation related to LDC's preservation practices. Overall, LDC's preservation plan demonstrates a commitment to the long-term accessibility and integrity of the data within its repository, employing established best practices and utilizing a robust infrastructure to ensure ongoing preservation efforts.

Report 7: Roper Center for Public Opinion Research

The Roper Center for Public Opinion Research is a sustainable domain repository hosted by Cornell University. The selection criteria for this repository analysis was conducted mainly due to the fact that this repository is owned by an academic institution. It has a long history of managing and providing access to public opinion data. Its mission is to collect, preserve, and disseminate public opinion data while improving the practice of survey research. The Roper Center addresses the criterion through its robust data curation and processing procedures. It assesses the content, structure, and format of submitted data, enhances it with metadata, and ensures the protection of

respondent confidentiality through de-identification and disclosure control methods. Documentation is compiled to provide users with necessary information.The Roper Center demonstrates a high level of compliance by adhering to its digital preservation policy, engaging in national digital preservation consortia, and requiring data providers and member institutions to sign agreements. These agreements grant the Center rights to archive and disseminate the data while outlining obligations and distribution information. Compliance levels may vary depending on the adherence of data providers, institutions, and users to the agreed-upon terms and conditions. The Center ensures confidentiality and ethics by removing identifying information from submitted data and employing disclosure control measures. These statements are supported by various pieces of evidence. The provided links offer access to official documents, such as data provider agreements, data deposit forms, membership agreements, and terms and conditions of use. Additionally, the description of procedures, practices, and participation in digital preservation consortia serves as evidence of the Center's commitment to preservation and compliance. Overall, the Roper Center for Public Opinion Research demonstrates a strong commitment to data preservation, ethical practices, and compliance through its comprehensive procedures, partnerships, and documentation.

Preservation Plan: It seems that Roper Center for Public Opinion Research has a comprehensive Digital Preservation Policy that guides its long-term preservation planning for digital assets. The policy acknowledges the challenges posed by changing technology and user needs. It emphasizes normalization and migration procedures as primary strategies to address file format obsolescence. It is described that the Center actively monitors file format obsolescence through its participation in the Data Preservation Alliance for the Social Sciences (Data-PASS) technology responsiveness program. The Digital Preservation Policy tries to ensure actions necessary for long-term usability are incorporated throughout the data curation workflow. The Center employs robust metadata management strategies to make data readable, meaningful, and independently understandable in perpetuity. However, it acknowledges that no assurance can be made regarding the complete effectiveness of these measures. The Center's participation in the Data-PASS consortium further strengthens its commitment to long-term preservation, even in situations where it may no longer retain archived material. The data provider agreement, cosigned by the Roper Center and its data providers, grants the Center the authority and rights to fulfill the obligations of its preservation policy, including transforming, storing, preserving, and disseminating the data. In conclusion, the Roper Center demonstrates a proactive approach to digital preservation by addressing format obsolescence, incorporating preservation measures in its workflows, and leveraging partnerships and agreements to ensure long-term accessibility and usability of its digital assets.

Report 8: Polar Data Catalogue, University of Waterloo

Created as a collaboration between departments of the Canadian government, the University of Waterloo and Neotix research Inc., the main objectives of the CCIN/PDC have been to provide data and information management infrastructure for the Canadian Cryospheric research community. We selected this repository as one which is owned and maintained by a University, which in this case is described to be the University of Waterloo. The Polar Data Catalogue (PDC) addresses each criterion through various strategies and measures. They have developed partnerships and collaborations with organizations such as ArcticNet, Indigenous and Northern Affairs Canada, the Department of Fisheries and Oceans Canada, and the Canadian Cryospheric Information Network (CCIN) to facilitate information exchange and data management. These partnerships provide funding, guidance, and data contributions to ensure new and relevant research is made available to the public. The PDC also advocates for free and open access to data and has implemented a data policy promoting data exchange and accessibility. They have established protocols for handling sensitive data, allowing submitters to classify their documents into limited categories to restrict access. The PDC is committed to preserving data integrity, availability, and accessibility, as well as ensuring confidentiality and privacy through their privacy policy and information security measures. The self-assessment does not explicitly mention different levels of compliance or provide information about variations in compliance levels for each criterion. However, the overall tone of the description suggests a strong commitment to meeting the requirements across all criteria. The PDC demonstrates a comprehensive approach to data management, preservation, accessibility, and security, indicating a high level of compliance with the specified criteria. The PDC provides evidence to support its statements through references to official documents, policies, and guidelines. They provide links to resources such as the Data Policy, which promotes free data exchange, and the privacy policy, which outlines their commitment to preserving user privacy. They also mention partnerships and affiliations with various organizations and institutions, showcasing their collaboration and contribution to data preservation and accessibility. While the provided information focuses on positive examples of compliance and evidence, no negative examples or shortcomings are mentioned in the self-assessment. The emphasis is on highlighting the PDC's efforts, policies, and partnerships to demonstrate compliance and commitment to the specified requirements.

Preservation Plan: It is described that the preservation plan for the Polar Data Catalogue (PDC) encompasses physical preservation and long-term accessibility. In terms of physical preservation, the PDC outlines its liability and warranties in its Terms of Use document. In the event of involuntary organizational dissolution, the PDC has a plan in place for the long-term preservation of its holdings. This plan involves exporting all organizational assets, including metadata, data, database, and VM configurations. Additionally, the PDC has a dissemination plan to ensure the continued accessibility of the preserved data. Regarding long-term accessibility, the PDC is said to have started taking steps to address the FAIR principles (Findable, Accessible, Interoperable, and Reusable). This involves implementing technical solutions in collaboration with the Canadian Consortium for Arctic Data Interoperability (CCADI). The PDC is reevaluating its internal data standards to ensure the perpetuation of file formats and standards. They recognize that certain segments of their data holdings may require intervention for proper preservation. Incorporating their recent emphasis on FAIR principles, the PDC intends to update its mission statement to accurately reflect its latest views and objectives regarding the preservation and dissemination of polar data. Overall, the preservation plan demonstrates the PDC's commitment to physical preservation and long-term accessibility. They are actively addressing the challenges of file format preservation and incorporating FAIR principles to ensure the enduring availability and usability of their data.

Report 9: The Odum Institute for Research in Social Science, University of North Carolina at Chapel Hill

H. W. Odum Institute for Research in Social Science ("Odum Institute") at the University of North Carolina at Chapel Hill ("UNC"), operates the Odum Archive, established in 1969, which is focused on the long-term preservation, access, and reuse of social science research data and associated materials. This repository may seem to be different than the rest of the selected ones that are being analyzed in this report, the reason we decided to review it was mainly because the report is described to be containing one of the largest catalogs of machine-readable social science data in the United States, encompassing collections like the Louis Harris Data Center, the National Network of State Polls, the Carolina Poll, and the comprehensive assortment of 1970 United States Census datasets. The Odum Archive addresses each criterion by implementing specific policies, procedures, and workflows. They have established clear guidelines and documentation to ensure compliance with each requirement. For example, they have a Digital Preservation Policy, Collection Development Policy, Data Curation Workflow, and Standard Operating Procedures for Data Deposit. These documents outline the steps and processes involved in preserving, managing, and distributing digital data assets. They also have a Data Deposit Form that depositors must complete and sign, granting necessary permissions for the use and preservation of the data. This repository demonstrates different levels of compliance depending on the specific requirements. Generally, the differences between compliance levels are not drastic, indicating a consistent adherence to the criteria. The repository tends to meet the requirements at a high level, with minor variations in the completeness or formalization of certain aspects. For instance, the review

mentions that the differences between compliance levels 3 and 4 are not significant, while the differences between levels 1 and 2 are more pronounced. However, in the case of the Odum Archive, there were no level 1 or 2 scores, indicating a strong overall compliance with the requirements. The Odum Archive provides comprehensive evidence to support their statements and demonstrate compliance. They refer to their official policies, procedures, and guidelines, such as the Digital Preservation Policy, Collection Development Policy, and Data Curation Workflow. These documents outline the repository's strategies, standards, and best practices. Additionally, the Odum Archive provides links to relevant resources and external documents that further support their compliance efforts, such as the Digital Curation and Preservation Framework. While the provided information focuses on positive examples of compliance, the absence of level 1 and 2 scores suggests that there were no significant shortcomings or negative examples identified during the review. The repository's detailed responses and clear policies indicate a thorough approach to meeting the requirements. The review specifically highlights the Odum Archive as a repository that exhibits a high level of detail, clear policies and workflows, and detailed evidence, indicating a positive assessment of their compliance efforts.

Preservation Plan: The Odum Archive Preservation Plan outlines a comprehensive and standards-based strategy for the long-term management and preservation of digital data collections. It seems to align with the Digital Curation and Preservation Framework's seven attributes of a trusted digital repository, including OAIS compliance, administrative responsibility, organizational viability, financial sustainability, technological and procedural suitability, systems security, and procedural accountability. The preservation plan is supported by the Odum Institute Data Archive Digital Preservation Policy, which establishes guidelines for preserving, managing, and distributing digital data assets and associated materials. Data selection and appraisal are based on the Odum Institute Data Archive Collection Development Policy, considering factors such as the collecting scope, quality of materials, and file formats. Depending on the value and processing requirements of the data, the repository applies three levels of curation, with a minimum requirement of preserving files at the bit-level and generating tabular files for preservation alongside the original files. It is described that the Odum Institute Data Curation Workflow guides the preservation actions taken during the four stages of the ingest process: deposit, triage, processing, and access. Standard Operating Procedures for Data Deposit ensure adherence to archival standards and best practices, including quality control inspections, file normalization, and metadata generation. Overall, it is obvious that the Odum Archive Preservation Plan demonstrates a comprehensive approach to digital preservation, encompassing policies, workflows, and standardized procedures to ensure the long-term accessibility and usability of the data collections it manages.

Report 10: CLARINO Bergen Centre, University of Bergen Library

The Clarino Bergen Centre is a repository which helps researchers in the fields of Humanities and the Cultural and Social Sciences, to manage, store, prepare, access, and analyze language related data. The CLARINO Bergen Centre is operated by the University of Bergen (UiB), which is of a similar profile as the majority of institutions we're analyzing in this report. As part of the Norwegian national CLARINO research infrastructure and follows the mission and goals of CLARIN ERIC, which aims to make language resources easily accessible for researchers and bring eScience to humanities disciplines. The repository addresses various self-assessment requirements in a detailed and thorough manner. They have established licenses (R2) for each resource, ensuring that submitters and users are bound by agreements. Submitters must be reliably authenticated, and the repository maintains a log of accepted licenses. Continuity of access (R3) is emphasized, with the repository being a permanent infrastructural offer, and the University of Bergen committing to maintaining its part of the CLARINO infrastructure until at least 2029. Confidentiality and ethics (R4) are taken seriously, with submitters accepting a distribution license agreement that confirms their responsibility and right to grant distribution rights. Authentication and logging mechanisms are in place to track submissions and resource access. Repository editors validate each submission, ensuring compliance with licenses, terms of service, privacy, and ethical considerations. Expert guidance (R6) is provided by a dedicated team at the University of Bergen Library, which collaborates with other departments and institutions. The repository's technical infrastructure follows the OAIS model, with clear procedures for data analysis, high-quality web interfaces, and regular reviews. Various measures, such as backups, database snapshots, and long-term storage in the DataVault, are implemented to ensure data safety and meet the requirements. Overall, the CLARINO Bergen Centre demonstrates a high level of detail, clear policies, and thorough evidence to meet the self-assessment requirements. The repository's commitment to accessibility, continuity, confidentiality, and expert guidance positions it as a reliable resource for researchers in the language-related disciplines.

Preservation Plan: The repository's digital assets are considered part of the University of Bergen's (UiB) digital assets and will be long-term preserved accordingly. The cloud infrastructure used by the repository is interchangeable and they say can be replaced by other providers if necessary. The UiB IT department ensures data persistence and performs regular backups. The repository follows ISO/IEC 27001 standards for backup routines, security measures, and crisis management, which are covered by the University of Bergen IT department. The repository has the right to move data to new media and convert it to new formats for long-term preservation, although such conversions have not been necessary so far due to the use of widely accepted open formats. The repository logs relevant actions automatically, and its strong link to CLARIN and adherence to open data, open standards, and open software contribute to the longevity of data access. Participation in CLARIN committees and working groups is part of the preservation plan to maintain the repository's connection to the broader CLARIN community. The UiB Library is committed to investing the necessary efforts to sustain long-term access to and preservation of the data in the repository.

In summary, the CLARINO Bergen Centre repository has a comprehensive preservation plan that involves the University of Bergen's IT department, adherence to ISO/IEC 27001 standards, Distribution License Agreements for each item, the use of open formats, and collaboration with the CLARIN community. The repository's commitment to long-term access and preservation is documented in its preservation policy.

TU Wien Research Data (TUWRD) Review

Preservation Plan: TU Wien Research Data is an institutional repository for only research data that is open to all users of TU Wien, including students [1]. This means that while students are subject to additional checks, other users who are outside of TU Wien must have their uploads approved by someone working at TU Wien (someone with an employment contract) [1]. Additionally, TUWRD allows users to select the license with which to use their data. TUWRD allows anyone to use the data (assuming it is not uploaded as restricted, hidden, or embargoed), so all data uploads with type open do not have to be a member of TU Wien to use the research data [1]. TUWRD does offer restricted and invisible to public upload policies for sensitive data. Removal of data which violates the policies of the TUWRD will still lead to the Personal Identifier (PID) being retained, as well as a tombstone citation so as not to result in broken links, and to retain the URL of the original record [1]. For the longevity of data, TUWRD offers storage of data for the life of the TUWRD, and notes that in the event that TUWRD shuts down, the data stored within will be transferred to another repository [1]. As TUWRD is operated jointly by TU Wien and the Center for Research Data Management, its operation is secured long-term [1]. All formats are allowed, however there is no guarantee for proprietary formats that preservation is guaranteed, due to the nature of proprietary formats [1]. There are two replicas of the repository which can be restored from, although the details of where these are stored are not mentioned [1]. This also means that

TUWRD protects the bit-level preservation of all documents contained in the repository as well, and the repository will try to provide readability and accessibility, along with migration to new formats as necessary [1]. Where new formats cannot be used, emulation will be used to try to maintain readability and accessibility [1]. All files have checksums which are checked periodically to ensure no changes occur to the stored data [1].

As a reviewer we would rate this repository a 3, as a formal agreement for which repository the data will be transferred to is not clearly agreed, and the periodicity of the checksum checks is also not clear. However, the repository is clearly taking the preservation of the data seriously and has numerous backups, ways to prevent the breaking of links, and appears to conform to FAIR principles. Overall, the differences between levels 3 and 4 tend to not be large, and in this case because of a small number of shortcomings/lack of public information, we feel it necessary to give it a 3.

[1] https://www.tuwien.at/index.php?eID=dms&s=4&path=Directives%20and%20Regulations%20of%20the%20Rectorate/Research_Data_Deposit_Policy.pdf