Profiling Open Digital Repositories in the Atmospheric and Climate Sciences: An Initial Survey

Chung-Yi Hou, Cheryl A. Thompson, & Carole L. Palmer
Center for Informatics Research in Science & Scholarship
Graduate School of Library and Information Science
University of Illinois at Urbana-Champaign
Champaign, IL 61821
hou, cathmps2, & clpalmer @illinois.edu

ABSTRACT
As digital research data proliferate and expectations for sharing data escalate, the landscape of open data repositories is becoming more complex. Data sharing is especially important in the atmospheric and climate sciences, where data integration and reuse can have significant impact on the advancement of research on climate change. This study analyzed 38 open digital repositories in the atmospheric and climate sciences using 55 criteria to characterize the repositories based on content analysis of the repository websites. The resulting repository profiles offer systematic descriptions of capabilities, services, and other characteristics for identifying specific repository features and comparing them across repositories. The profiles are of value for data curation education, researchers in the geosciences, and librarians and others providing data services in the sciences.

Keywords
Data repositories, data sharing, digital technology, cyberinfrastructure, communication in science.

INTRODUCTION
Managing digital data has become a challenge for researchers and research organizations, especially with the rapid increase in the volume and types of data generated by new scientific instrumentation. At the same time, digital data are now understood to be important assets with great potential for sharing and reuse for new applications. Data repositories are key to realizing this potential as they serve to collect, organize, preserve, and provide access to large bodies of data, which can be combined and reused in innovative ways. Not surprisingly, to accommodate the abundance of data products, data repositories are proliferating at a high rate. For example, 977 data repositories are currently listed by DataBib, and this number will increase considerably when the resource merges with re3data.org under the auspices of DataCite to create a “single, sustainable registry of research data repositories” (re3data.org, 2014).

Thus, researchers are faced with navigating an increasingly complex landscape of repository options for depositing their data, to meet the expectations of their funders and publishers, and to adhere to more general national expectations for open data (Stebbins, 2013). Academic libraries that provide data consulting and referral services for their researchers must also maintain broad and detailed knowledge of the range of data repository options as well as the capabilities and services that the repositories provide.

This study analyzed data repositories in the atmospheric and climate sciences as part of the Data Curation Education in Research Centers (DCERC) project. DCERC is developing a sustainable and transferable model of data curation education for master’s and doctoral students in Library and Information Science (LIS). The project is led by the University of Illinois in collaboration with the University of Tennessee and the National Center for Atmospheric Research (NCAR), a premier national research center with state-of-the-art data operations and services. A key part of the program has been summer internships for master’s students at NCAR. Formative evaluation of the internship program in the first years pointed to a need for students to have better grounding in trends specific to data management in the geosciences in addition to the foundational academic work in data curation.

This analysis was conducted to develop a resource for data curation students to build more comprehensive awareness of data repository services in the atmospheric and climate sciences prior to their work at NCAR. The resource is also being optimized for data professionals working in the geosciences and for data science curriculum more
generally. The analysis characterizes 38 digital repositories in the atmospheric and climate sciences according to 55 criteria. The results provide systematic descriptions of criteria related to content, capabilities, services, and other characteristics of value for identification of specific features of a repository and comparison across repositories.

BACKGROUND AND RATIONALE
Many papers and reports discuss technical aspects of digital repository implementations (Bach et al., 2012; Barros, Laender, Gonçalves, Cota, & Barbosa, 2007), but fundamentally, digital repositories should also share a common set of goals that align with the values of the scientific community and other stakeholders. According to Lynch (2003), repositories should commit to data stewardship for the long term and provide a set of services to their community that enables data access, distribution, and collaboration. In addition, Case and Matz (2003) stress that repositories should allow “open access” for disseminating and using information in a cost-effective manner. It follows that an open digital repository needs to facilitate the combination of heterogeneous data types across science domains, distance, time, and scales. In the Earth and environmental sciences, for example, DataOne (Michener et al., 2012) and NEON (Keller, Schimel, Hargrove, & Hoffman, 2008) are currently implementing approaches to make progress to this end.

The approach used in “Scientific Data Repositories on the Web: An Initial Survey”, conducted by Marcial and Hemminger (2010), served as an important model for this study. Repository websites function as a critical point of discovery and access for data. They are the primary interface between a repository and their user communities, and the place where users will look for information related capabilities and services. Marcial and Hemminger identified key categories of information needed by data users and developed a framework for evaluating repositories’ success in delivering this information. Our study utilized their framework as a starting point for profiling atmospheric and climate science repositories.

RESEARCH OBJECTIVES
The profiling study of the open digital repositories was guided by the following objectives:

- Identify established repositories in the atmospheric and climate sciences.
- Inventory criteria related to content, capabilities, and services identified through content analysis of repository websites.
- Determine patterns and trends across repositories.
- Assess availability and accessibility of information provided by repositories.

METHOD
The team began with Marcial and Hemminger (2010) to determine the initial set of repository criteria to be profiled. This set of criteria was supplemented by consulting the Digital Curation Centre’s (DCC) Curation Lifecycle Model (http://www.dcc.ac.uk/resources/curation-lifecycle-model) and the Open Archive Information System (OAIS) Reference Model (http://public.ccsds.org/publications/archive/650x0b1.PDF) for additional important repository characteristics and functions. Criteria from both models were added to create the final list with corresponding definitions for systematic analysis across repository websites. Table 1 provides a sample of the criteria and the corresponding definitions.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>File Format/Type Accepted</td>
<td>What file formats does the repository accept on submission?</td>
</tr>
<tr>
<td>File Format Disseminated</td>
<td>Does the repository provide archived data digitally, physically or both? What are the formats?</td>
</tr>
<tr>
<td>Mission Statement</td>
<td>The mission statement or goals of the repository.</td>
</tr>
<tr>
<td>Producers</td>
<td>Persons, clients, or systems that provide the information to be preserved.</td>
</tr>
<tr>
<td>Consumers</td>
<td>Users that interact with repository to find and acquire archived data.</td>
</tr>
<tr>
<td>Long Term Sustainability</td>
<td>What is the information that the repository has defined for the long term sustainability of its collection?</td>
</tr>
<tr>
<td>Metadata Standard</td>
<td>Is the repository in compliance with any metadata standard?</td>
</tr>
<tr>
<td>Transfer of Rights</td>
<td>Are the rights to data transferred to the repository at the time of data submission? What are other legal details that occur during the submission process?</td>
</tr>
<tr>
<td>Data Citation</td>
<td>Does the repository provide a method for citing the archived data? If yes, what is the identified standard used?</td>
</tr>
</tbody>
</table>

Table 1. Sample of criteria and definitions.

The primary resources for selecting repositories for analysis were Databib.org (http://databib.org/index_subjects.php#Geo) and EarthCube (http://connections.earthcube.org/EcResources/Resources). Specifically, repositories tagged as “Atmosphere,” “Climatology,” and/or “Earth Sciences” for Databib.org and “Atmosphere” and/or “Climatology” for EarthCube were included in the selection.

An initial test was conducted to assess the process of extracting information about the criteria from the repositories’ websites. The team evaluated the information collection process for four repositories by reviewing all the webpages and links associated with the websites. Based on the test, the team refined and modified the list of criteria,
clarified the definitions, and removed those determined to be outside the scope of the study. For the full process, 38 repositories were identified. The website and links for each repository were examined to extract information for 55 criteria.

It is important to note that it should not be assumed that capabilities, services, and other aspects of a repository are uniformly covered by the repository’s website. Some of the missing criteria might be in practice within the repository but are not represented or discovered in our analysis of the content. A survey would be an important next step in validating these results, as was done in the Marcial and Hemminger (2010) study.

RESULTS

Overall, all websites contained valuable information, including management structure, funding sources, contact information, mission statements or goals, and data distribution model (distributed or centralized). Majority of the websites also provided information regarding the data formats accepted and disseminated, the data community in terms of data producers and consumers, data processing procedures, terms and conditions for data usage, data services and support, and data identification and citation methods. In addition, information with respect to appraisal and selection policy, data submission method, authorization and authentication of users, and research involvement were often included as well. However, coverage was surprisingly uneven for usage of standardized metadata schemes, transfer of rights, and data security. Information was also rare on certification and accreditation of the repositories, preservation policy, long term sustainability plans, and measures against physical damages or losses of data.

The results provided here related primarily to the key elements of the DCC’s curation lifecycle model. Table 2 provides a summary of criteria in relation to the curation lifecycle model.

All the repositories concentrated on data in digital formats; however, 18.4% (7) also accepted physical forms, such as microfilms, CD, and hard drives.

While 21.1% (8) of the repositories identified and followed specific metadata standards, such as ISO 19115 and Dublin Core, more than half utilizing a customized metadata scheme. While two repositories (5.3%) invited anyone from their data communities to provide and improve metadata content and quality, others depended mainly on data providers to contribute descriptive metadata at the time of data submission.

In terms of preservation planning, only five of the repositories (13.2%) had explicit preservation policies, and only 10.5% (4) provided evidence of complying with an identified certification or accreditation standard (e.g. Information Fair Trade Scheme, ISO 9001, and ISO 14001).

<table>
<thead>
<tr>
<th>Elements in DCC Curation Lifecycle Model</th>
<th>Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>File Format/Type Accepted, File Format Disseminated</td>
</tr>
<tr>
<td>Description and Representation Information</td>
<td>Metadata Standard</td>
</tr>
<tr>
<td>Preservation Planning</td>
<td>Preservation Policy, Certification and Accreditation</td>
</tr>
<tr>
<td>Community Watch and Participation</td>
<td>Data Services and Support</td>
</tr>
<tr>
<td>Create and Receive</td>
<td>Producers, Consumers</td>
</tr>
<tr>
<td>Appraise and Select</td>
<td>Appraisal and Selection Policy</td>
</tr>
<tr>
<td>Ingest</td>
<td>Submission Method, Transfer of Rights, Embargo Period</td>
</tr>
<tr>
<td>Preservation Action</td>
<td>Data Processing, Data Identification</td>
</tr>
<tr>
<td>Store</td>
<td>Long Term Sustainability, Protection Against Physical Damages/Losses</td>
</tr>
<tr>
<td>Access, Use, &amp; Reuse</td>
<td>Data Distribution Model, Access Authorization and Authentication, Fees, Data Citation</td>
</tr>
</tbody>
</table>

Table 2. DCC Curation Lifecycle Model elements and associated criteria.

All but one repository (2.6%) provided information on their services and support functions with 44.7% (17) providing software and tools. Most repositories (68.4%; 26) also provided support in the forms of outreach and education, presentations and seminars, and customer service.

Most repositories (60.5%; 23) primarily acquired data through deposit by the members of their data communities, and 21.1% (8) of the repositories also received data directly from instruments. Although researchers were the main data producers for the repositories, only 10.5% (4) identified researchers as their only consumers. The remaining considered the general public as part of their user base, as well as the education, commercial, and policy sectors.

Although only 15.8% (6) of the repositories had specific data appraisal and selection procedures available, 71.1% (27) indicated preferences for certain data types. For example, the selection of data might depend on criteria, such as types of instruments, research topics, or science disciplines that would help complement and enrich the repositories’ existing data collection. In addition, 31.6% (12) required data providers to contact the repositories to arrange the details for data submissions.

During the ingest process, 31.6% (12) of the repositories addressed issues regarding transfer of rights, but only 7.9% (3) of the repositories specifically stated that their datasets were “free from copyright.” Another 10.5% (4) of the
repositories would allow embargo periods while only 10.5% (4) of the repositories addressed policies regarding methods to process data that might have confidentiality concerns.

Nearly half of repositories (47.4%; 18) provided some level of data processing, and importantly, more than half of the repositories (52.6%; 20) used some type of data identification system. However, it appeared that only seven of these repositories (18.4%) used data identifiers consistently. Among the repositories that used data identification, DOI was the most commonly used scheme.

A long-term sustainability plan was evident for only one repository (2.6%), and only 15.8% (6) of the repositories covered measures for protecting against damages or losses due to environmental factors, such as natural disasters.

Repositories using a distributed access model (65.8%; 25) outnumbered those with a centralized model. Almost half of repositories (42.1%; 16) allowed access without authorization, such as requiring users to create registered accounts. Authorization was sometimes reserved for additional levels of data access, such as bulk downloading. One repository (2.6%) even required authentication and authorization in person. Depending on the data type, 23.7% (9) of the repositories might require fees under certain access and/or use conditions. A promising trend was seen in citation expectations, with 78.9% (30) of the repositories requesting that users cite and/or acknowledge the dataset and the data center for their contributions.

CONCLUSION
The above results show that repositories provide a significant amount of information to their data communities that are helpful in facilitating data sharing. They are particularly strong in providing information for the following criteria: available data format, services and support functions, data processing, access without requiring authorization, and methods for data citation and/or acknowledgement. However, aspects that are particularly important for building trust, such as compliance to accreditations and certifications and upholding confidentiality, are lacking for most. Moreover, it seems that access and reuse of data by all research communities can benefit from dedicated attention to guidelines and policies on metadata and transfer of rights. Policies for migration, reappraisal, and disposal are another area that is clearly underdeveloped.

Next steps in development of the profiles will be to expand the number of repositories covered and provide further cross-repository analysis of patterns and trends.

ACKNOWLEDGMENTS
This study was funded in part by the Data Curation in Research Centers (DCERC) project, through a grant from IMLS (RE-02-10-0004-10).

REFERENCES