Harvestable Metadata Services Development: Analysis of Use Cases from the World Data System

RESEARCH PAPER

ROBERT R. DOWNS
ALICIA URQUIDI DÍAZ
QI XU
JUANLE WANG
AUDE CHAMBODUT
CHUANG LIU
SIMON FLOWER
KAREN PAYNE

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

ubiquity press

ABSTRACT

Minimally, a research data repository exists to make a collection of data assets available to potential users. If a dataset cannot be discovered and found, it cannot be reused (Garnett et al. 2017). Harvestable metadata catalogues are a key strategy for achieving greater global findability of data assets, as they create a surveyable access point to discover data products within large data collections. Such catalogues can be especially effective if they are tailored for interoperability with feature-rich infrastructures (e.g. meta-catalogues, see Kapiszewski & Karcher 2020; CRFCB 2014) that are highly visible and widely used, and also themselves integrated within the larger ecosystem of research infrastructures.

This study offers insight into a set of World Data System (WDS) research data repositories ongoing and successful implementations of harvestable metadata services, which apply established and emerging research data standards and practices to fit global, local and domain-specific interoperability contexts. Establishing a harvestable metadata service involves making choices in a space where standards and technologies are continuously evolving. The repositories in this study leverage the resources they have, within the policy and funding constraints of their institution, to serve the changing needs of heterogeneous user groups. This document encapsulates and completes the work that was carried out by the WDS International Technology Office (ITO) Harvestable Metadata Services Working Group (HMetS-WG).

CORRESPONDING AUTHOR:

Dr. Robert R. Downs

Center for International Earth Science Information Network (CIESIN), Columbia Climate School, Columbia University, United States

rdowns@ciesin.columbia.edu

KEYWORDS:

metadata harvesting; metadata for discovery; World Data System; metadata catalogue; dataset findability

TO CITE THIS ARTICLE:

Downs, RR, Urquidi Díaz, A, Xu, Q, Wang, J, Chambodut, A, Liu, C, Flower, S and Payne, K. 2023. Harvestable Metadata Services Development: Analysis of Use Cases from the World Data System. *Data Science Journal*, 22: 20, pp. 1–20. DOI: https://doi.org/10.5334/dsj-2023-020

1. INTRODUCTION

Harvestable metadata services are an effective, established and widely-used approach to promoting data discovery and sharing across broad communities of potential data users, across multiple disciplines (Lokers et al. 2016; Valentine et al. 2020). For the purpose of this study, we understand harvestable metadata *as a set of metadata records in a standardized format and schema that is shared with aggregation services by means of specific protocols for metadata transfer, which are also standardized*. In this paper, we describe examples of ongoing and successful implementations of harvestable metadata services, which apply emerging and established standards and community practices to fit local and domain-specific research data management contexts. These use cases originated from the Harvestable Metadata Services Working Group (HMetS-WG),[1] which met frequently in a series of working sessions over 6 months during 2020, followed by occasional meetings during 2021. The study offers an overview of the infrastructures, standards and communities of the repositories that were members of the HMets-WG, as well as offering a wider-ranging discussion of challenges that repositories may face when developing data services, such as harvestable metadata.

Taking a qualitative approach, this study explores issues for implementing harvestable metadata services at repositories. We start with a description of use cases, focusing on each repository's technical features, along with the challenges encountered in pursuit of repository-defined and community-oriented service development goals. Repositories are also characterized by the subject and disciplinary areas covered, targeted user groups, and services offered. The full-length profiles for each repository are described as use cases by Urquidi Diaz et al. (2022). After examining the use cases within the context of the current literature on recommended practices for metadata syndication and pathways toward interoperability, we present a set of common characteristics and challenges described by the repositories in this study. These experiences involved making decisions about which technologies to develop for an often heterogeneous dynamic user base, within an evolving technological landscape, in order to implement data and metadata services that fit within the resource and policy constraints of the repository.

METADATA HARVESTING, STANDARDS AND PROTOCOLS

In a typical metadata sharing process, a research data repository will share a catalogue of assets: a collection of metadata records that describe each dataset, which are typically accessible through a search interface on the repository's portal. The repository may also share a set of standardized metadata records via additional access points (or harvestable metadata services), using a metadata transfer protocol through which aggregation services, such as harvesters, obtain the metadata (see Figure 1). Persistent links to data landing pages at the host repository are typically contained in those records. An aggregator may then convert (re-format or cross-walk) the acquired records into a unified display standard, to be disseminated by means of a federated metadata catalogue or a federated search engine. Examples of metadata harvesters that target research data include the Canadian Federated Research Data Repository (FRDR),[2] and B2FIND (Europe).[3]

The adherence to shared standards and community practices is a key tenet for successful digital research infrastructure (DRI) integration and interoperability (Dietze et al. 2018; Waide, Brunt & Servilla 2017; Yu et al. 2021). Common standards for harvestable metadata include the Dublin Core (DCMI 2020), DataCite (DataCite Metadata Working Group 2021) and ISO 19115 (International Standards Office 2019) metadata schemas, as well as protocols for transferring metadata, like the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) (Lagoze et al. 2005) and the Open Geospatial Consortium Catalogue Service for the Web (OGC-CSW) (Nebert, Voges & Bigagli 2016). Metadata records are usually transferred as eXtensible Markup Language (XML) or JavaScript Object Notation (JSON, and JSON-LD, for linking data within semantic metadata) files. Another approach to syndicating metadata uses semantic metadata tags, such as Schema.org,[4] that are placed in the HTML of dataset landing pages

1 The group was led by the International Technology Office[25] of the International Science Council's (ISC) World Data System (WDS). WDS-ITO, https://wds-ito.org.

2 Federated Research Data Repository (FRDR), https://www.frdr-dfdr.ca/repo/.

3 EUDAT Collaborative Data Infrastructure – B2FIND, https://eudat.eu/services/b2find.

4 Schema.org, https://schema.org/.

on a repository's web portal, or in separate metadata files. This strategy relies on web crawlers (such as Google) parsing the semantic metadata to aggregate and index the landing pages for search engine retrieval. Even though this approach to metadata sharing can complement harvestable metadata services, the semantic strategy was not pursued by the HMetS-WG. Instead, WDS-ITO engaged members of various communities in a separate initiative to develop semantic metadata using Schema.org (Payne & Verhey 2022).

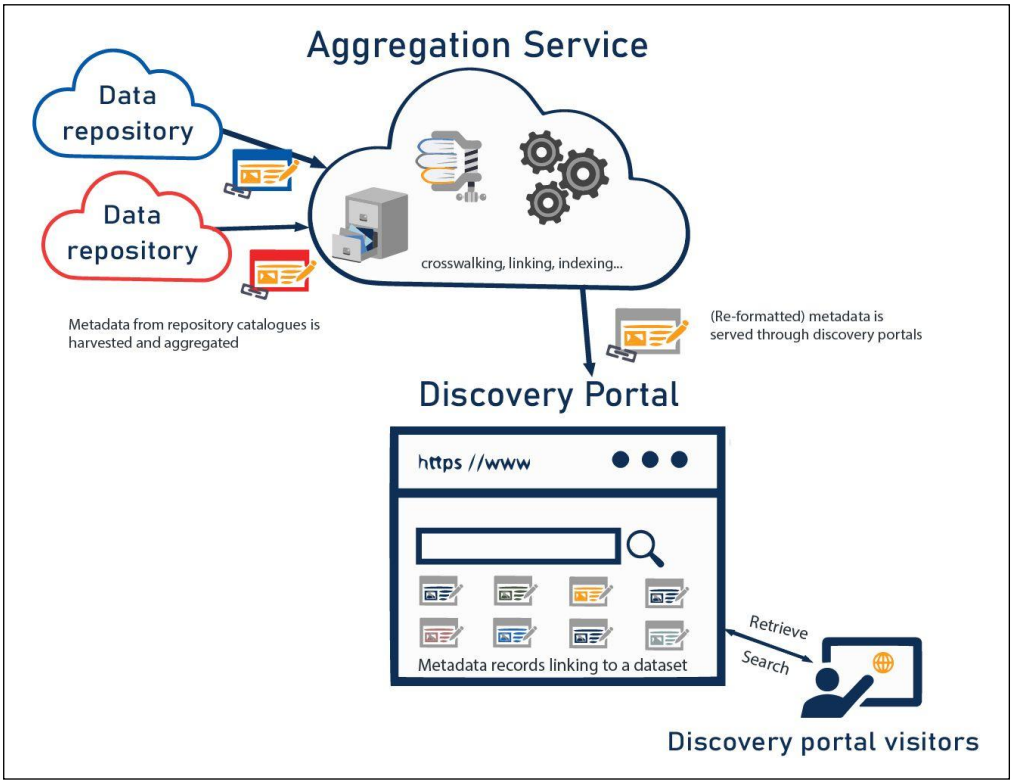

Figure 1 The metadata harvesting process. Standardized metadata is harvested from repository catalogues, then processed by an aggregation service. The service disseminates the metadata records through a search and discovery portal and/or by serving it to further aggregation services for distribution.

As the importance of reusing data is increasingly recognized across disciplines, data repositories have proliferated to meet this demand, with the number of data repositories listed in repository registries, such as the Registry of Research Data Repositories (Re3Data),[5] growing rapidly (Culina et al. 2018). Also, with each data repository offering additional open data products, finding a particular dataset of interest becomes challenging for potential data users (Kramer, Klas & Hausstein 2018; Plante et al. 2021). A recognized approach to address this challenge is for repositories to establish capabilities for harvesting metadata to facilitate searchability and global discoverability (Culina et al. 2018; Plante et al. 2021; Wu et al. 2019). Furthermore, the availability of harvestable metadata is an indicator of dataset findability (CoreTrustSeal 2022; FAIR Data Maturity Model WG 2020) and contributes to repository TRUST-worthiness as it improves integration with the wider data management community (Lin et al. 2020: 3).

2. RESEARCH METHODOLOGY

DATA COLLECTION

In September 2019, the WDS-ITO invited WDS member repositories to participate in the HMetSWG and to (optionally) serve as use cases for the study. The invitation was sent to 35 unique WDS member organizations that had previously expressed interest in being informed about new WDS initiatives. Nine WDS member repositories participated in the group (Table 1), and seven (Table 2) were adopted as use cases. Over the course of the group's sessions, the repositories presented an overview of their infrastructure, data holdings, and services. All participating repositories also provided the group with a schematic overview of their features and subsequently, three repositories (NSSDC, INTERMAGNET and SEDAC) also completed the implementation plan template, described below, and shared these with the WDS-ITO.

5 The Registry of Research Data Repositories, https://www.re3data.org/, is a global registry of research data repositories.

The group's work agenda was initially guided by a workflow structure proposed by WDS-ITO (Figure 2), which represents harvestable metadata services development as a set of discrete, successive steps. As group discussions progressed, WDS-ITO provided members with a Harvestable Metadata Services Implementation Plan template (Urquidi Diaz 2021b) to describe their implementation plans. The template was inspired by and borrowed heavily from the CESSDA-Saw guidance package (Bornatici et al. 2017) and the JISC project plan templates (JISC 2011) for data service planning, which were designed to be adapted to specific use cases for a single service or a subset of services, such as harvestable metadata services. Both of these resources also include guidance for drafting implementation plans for these types of services (Bornatici et al. 2017; JISC 2011), which also informed the development of the template. Supporting information resources also included a Twine interactive narrative/storyfied walk-through of the implementation plan flowchart (Urquidi Diaz, Li & Payne 2021), and a Zotero library with resources related to harvestable metadata services (Urquidi Diaz 2021a). While the questions derived from the workflow structure guided initial HMets-WG discussions, the availability of these additional resources, along with the individual repository overviews and implementation plans, facilitated broader discussions of implementation issues among the HMetS-WG repositories.

Table 1 HMetS-WG participants, with WDS membership type and host institutions.

WDS MEMBER	TYPE	HOST INSTITUTION(S)
Centre de Données Astronomiques de Strasbourg (CDS)	Regular	Strasbourg Astronomical Observatory (ObAS); University of Strasbourg; French National Centre for Scientific Research (CNRS)
Global Change Research Data Publishing and Repository (GCdataPR)	Regular	Institute of Geographical Sciences and Natural Resources Research (IGSNRR), Chinese Academy of Sciences (CAS); Geographical Society of China
International Real-time Magnetic Observatory Network (INTERMAGNET)	Network	Multiple institutions (worldwide)
International Service of Geomagnetic Indices (ISGI)	Regular	School and Observatory of Earth Sciences (EOST); University of Strasbourg; French National Centre for Scientific Research (CNRS)
International GNSS Service (IGS)	Network	Multiple institutions
National Space Science Data Center (NSSDC)	Regular	National Space Science Center (NSSC), Chinese Academy of Sciences (CAS)
Socioeconomic Data and Applications Center (SEDAC)	Regular	Center for International Earth Science Information Network (CIESIN), Columbia University; Earth Observing System Data and Information System (EOSDIS), National Aeronautics and Space Administration (NASA)
World Data Center for Geomagnetism (Edinburgh)	Regular	British Geological Survey (BGS)
World Data Centre for Renewable Resources and Environment (WDC-RRE)	Regular	IGSNRR; CAS

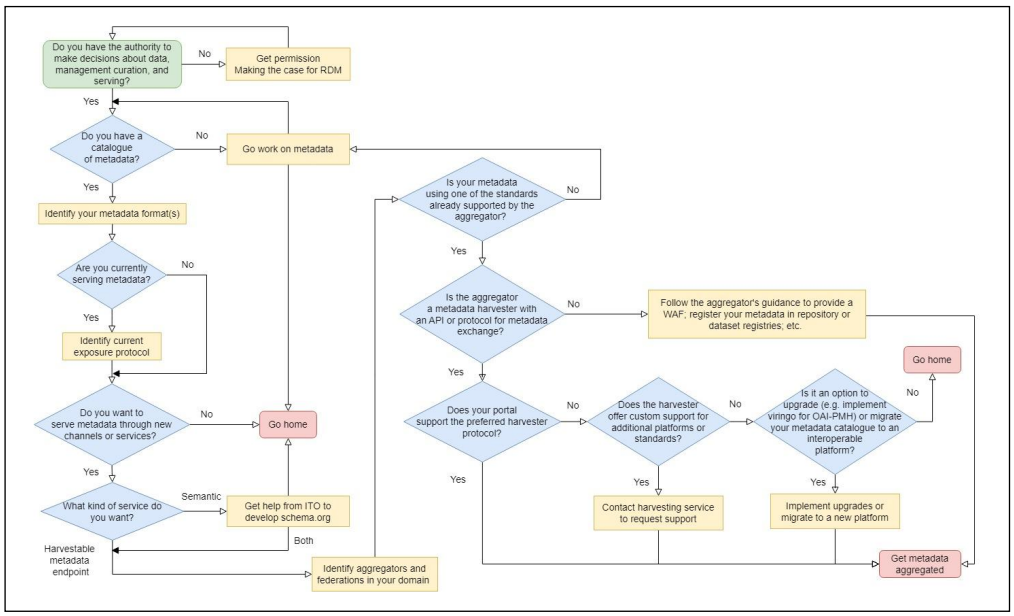

Figure 2 Flow-chart diagram of a typical harvestable metadata services implementation (Payne, Urquidi Diaz & Li 2021). This diagram gives a schematic representation of the steps involved in creating a harvestable metadata service. The HMetS-WG used these steps to scaffold the group's initial work.

DATA PROCESSING AND ANALYSIS

Building on the initial discussions of the workflow questions, the subsequent broader discussions among the HMets-WG repositories further contributed to the development of detailed repository profiles, which are accessible online as use cases (Urquidi Diaz et al. 2022). Where available, the profiles reference the repositories' technical documentation and other relevant publications to provide informative use cases. Urquidi Diaz et al. (2022) described the following characteristics of the repositories within the use cases:

1. Institutional overview: Brief description of the repository's institutional context: Its governance, history, mandate, mission, memberships, and other organizational features.
2. User community: Target communities for repository services.
3. Infrastructure overview: Description of repository's data holdings and technical infrastructure for service provision.
4. Current state of metadata: Metadata formats, standards used, and metadata services (if any).
5. Planned development: Plans for future development.
6. Resources: Description of repositories' sources of support and financing.
7. Challenges: Initially, each repository described the challenges they have faced in developing harvestable metadata and other data services on their platforms.

Discussing the institutional overviews and implementation plans, as well as the compiled information resources, in terms of applicability to repository practices, contributed to understanding the current state of the repositories implementation issues. While differences across the repositories were observed, discussions about the common challenges that the repositories faced when considering the issues associated with the development of harvestable metadata services identified similarities among the challenges of the repositories represented in the HMetS-WG. Recognition of these similarities led to the emergence of a consensus on the challenges that the participating repositories face for the development and deployment of harvestable metadata services.

3. THE HMETS-WG SET OF USE CASES

3.1. REPOSITORIES PARTICIPATING IN THE WORKING GROUP

The host institutions of the participating repositories (see Table 1) were based in China, the UK, the US, and France. Seven repositories were Regular[6] Members of the WDS and two were Network[7] Members (WDS Scientific Committee 2016).

3.2. RESEARCH AREAS AND TARGET USER COMMUNITIES

The research areas served by the repositories represent a predominant Earth- and planetary sciences orientation. Social sciences, including environmental and economic sciences, also are strongly represented. As described by Urquidi Diaz et al. (2022), three repositories can be classified broadly as social and environmental science research centers that focus on spatial data: The World Data Centre for Renewable Resources and the Environment (WDC-RRE),[8] Global Change Data Publishing & Repository (GCdataPR),[9] and the Socioeconomic Data and Applications Center (SEDAC).[10] Two repositories, the Chinese National Space Science Data Centre (NSSDC)[11]

6 '[O]rganizations that are data stewards and/or data analysis services (e.g., data centres and services that support scientific research by holding and providing data or data products)' (WDS Scientific Committee 2016).

7 '[U]mbrella bodies representing groups of data stewardship organizations and/or data analysis services, some of which may or may not be WDS Regular Members who usually serve as coordinating agents for nodes that have common characteristics and mostly common disciplines' (WDS Scientific Committee 2016).

8 World Data Centre for Renewable Resources and Environment, http://wdcrre.data.ac.cn/.

9 Global Change Research Data Publishing and Repository, http://www.geodoi.ac.cn/.

10 Socioeconomic Data and Applications Center, https://sedac.ciesin.columbia.edu/.

11 National Space Science Data Center of China, https://www.nssdc.ac.cn/eng/.

and the International GNSS Service (IGS),[12] can be categorized as representing astronomy and geodesy (Urquidi Diaz et al. 2022). Lastly, the International Real-time Magnetic Observatory Network (INTERMAGNET)[13] and the International Service of Geomagnetic Indices (ISGI)[14] are dedicated to managing and sharing geomagnetic research data and related data products (Urquidi Diaz et al. 2022).

REPOSITORY	SUBJECT AREAS	USER GROUPS
GCdataPR	Agriculture, Area studies, Earth sciences, Economics, Environmental studies, forestry, Geo-ecosystems Geography, and History	Global change students, researchers policy makers and society in China and worldwide
IGS	Earth sciences, Geodesy, GNSS, GPS, Precise positioning, Navigation, Timing, and Space sciences	Mainly IGS staff, project and working group participants. More broadly: worldwide users of modern mapping, orientation and navigation systems, enterprises, non-profits, institutions and government actors
INTERMAGNET	Earth sciences, Geomagnetism, Space sciences	Scientific community, geomagnetism community, members of IAGA,[14,15] commercial users
ISGI	Solar-Terrestrial physics, Space weather-Space Climate, Space sciences, Earth sciences, Geomagnetism	Academia (including behavioral biology), members of IAGA communities, private and public sectors (military, telecommunications, satellite operators)
NSSDC	Astronomy, Computer sciences, Planetary science, Space physics, Space sciences, Space weather	Typical users are Chinese and international researchers in subject areas
SEDAC	Agriculture, Architecture and design, Anthropology, Area studies, Business, Chemistry, Climate science, Computer sciences, Cultural and ethnic studies, Earth sciences, Economics, Engineering, Environmental science, Environmental and forestry studies, Geography, Health sciences, Information system science, Political science, Sociology, Statistics, Sustainability science, Systems science, Transportation	User community interested in studying human interactions in the environment
WDC-RRE	Earth sciences, Ecology, Environmental studies and forestry, Geography, Geoinformatics, Natural resources	Mainly academic researchers and students, also scientific staff and technicians, general public, government agencies, policy makers, and international organizations

Table 2 Subject areas represented by repositories and target users groups. Subject areas were provided to WDS-ITO by the repositories.

3.3. REPOSITORY FEATURES

Table 3 gives an overview of each repository's technical features: the type of repository platform and catalogue service used, metadata standards and protocols, and a list of any current, known aggregators of their metadata assets. Figure 3 presents the metadata exchange protocols utilized by the repositories studied, in the context of those of the larger WDS membership, as surveyed in 2019 by the WDS-ITO (Payne & Urquidi Diaz 2020). Relative to WDS members previously surveyed, the repositories in the use cases have, or plan to develop, more OGC-CSW and Opensearch, and fewer OAIPMH services (Urquidi Diaz et al. 2022). It also should be noted that the WDS member survey data reported by Payne and Urquidi Diaz (2022) does not distinguish between protocols residing within repositories and those that are provided by aggregators, such as the Earth Observing System Data and Information System (EOSDIS) and the Global Earth Observation System of Systems (GEOSS), that disseminate metadata on behalf of repositories.

3.3.1. Participation in research data networks

As described in the sections below, it appears that participation in national, regional, as well as subject-specific networks has generally shaped the repositories' infrastructure, particularly in the ways that their adoption of harvestable metadata services has developed or is being planned for development.

12 International GNSS Service, https://igs.org/.

13 International Real-time Magnetic Observatory Network, https://intermagnet.github.io/.

14 International Service of Geomagnetic Indices, https://isgi.unistra.fr/.

15 International Association of Geomagnetism and Aeronomy, http://www.iaga-aiga.org/.

REPOSITORY	REPOSITORY PLATFORM & CATALOGUE	METADATA STANDARDS	METADATA SERVICE PROTOCOLS	KNOWN AGGREGATORS
GCdataPR	Custom GCdataPR 2.0	DCI[16], DataCite	OpenSearch	CrossRef, China-GEOSS, CNKI, DCI, CSTR, ScienceEngine
IGS	Catalogue via NASA CMR – *Developing new discovery platform*	DIF 10, ECHO 10, ISO 19115-2:2009 (MENDS and SMAP dialects), UMM-C	CMR CSW, CMR public APIs, OpenSearch	via NASA's CMR
INTERMAGNET	Custom repository, with some datasets on GFZ Potsdam data repository	Via INTERMAGNET: IAGA2002, CDF; Via GFZ: GeoJSON, DataCite, ISO 19115	Via homepage: HTTP, FTP; Via GFZ: request to DataCite's API	DataCite, FIDGEO[17]
ISGI	Custom – *Public access metadata service*	IAGA2002 – *CERIF, DataCite, and/or DCAT based profiles and/or crosswalks*	Via homepage: HTTPS; request to DataCite's API	
NSSDC	Custom	NSSDC Core Metadata Specification, SPASE – *DataCite, Data model compatible with NSSDC*	OpenSearch, OGC-CSW (via WDS China), Data search platform, – *OAI-PMH*	National Science and Technology Data Sharing Network of China, Scientific Data Center, CAS
SEDAC	Vital Digital Asset Mgt. System (Fedora) – *Migrating to Drupal 8*	FGDC CSDGM, ISO 19115, DataCite	IDN OGC CSW, NASA CMR CSW, CMR public APIs, OpenSearch	DataCite, GEOSS (via EOSDIS/ CMR)
WDC-RRE	Custom: Debian OS, OSS NGNIX, PostgreSQL, TorCMS	Dublin Core, ISO 19115, custom Data Identification and Metadata Standards – *Revision planned*	OpenSearch, OGC-CSW 3.0.0, OAI-PMH 2.0, SRU 1.1., – *Geonetwork*	WDS-China, CNKI

Table 3 Use Case Infrastructures: Summary of Features.

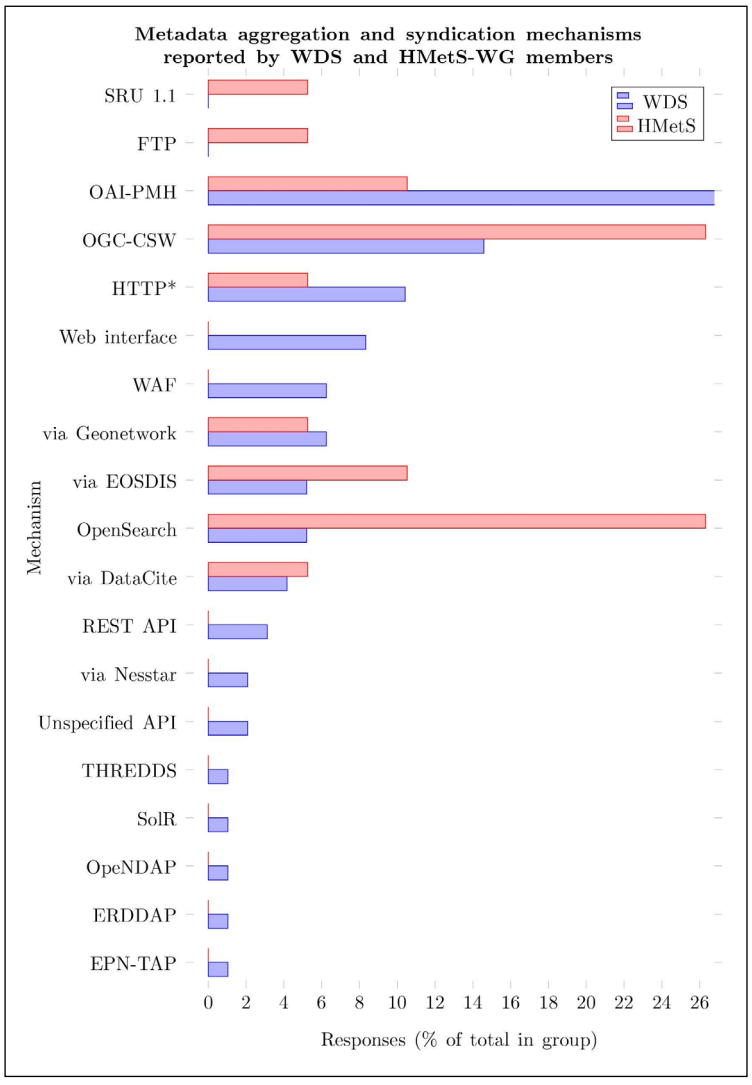

Figure 3 This bar chart compares the mechanisms for metadata exposure (aggregation, discovery, etc.) that were reported by the HMetS-WG repositories with those reported by the WDS repositories in a 2019 member survey (Payne & Urquidi Diaz 2020: 11, 15). Since some repositories reported serving their metadata via third-party services, these services also have been included (e.g. DataCite, EOSDIS, etc.). *Includes schema.org.

16 Data Citation Index, https://clarivate.com/webofsciencegroup/solutions/webofscience-data-citation-index/.

17 Specialized Information Service for Geosciences, https://www.fidgeo.de/en/fid-geo-en/.

All of the studied repositories have been guided or supported by a larger entity while developing harvestable metadata services: INTERMAGNET and ISGI have participated in the European Open Science Cloud's (EOSC) EPOS ERIC project, while WDC-RRE, NSSDC and GCdataPR have developed with support from Chinese research data institutions. One of the data sources for the GCdataPR comes from cooperation with journals for enabling discovery. GCdataPR initiated a tri-journal program since 2015 to facilitate dataset publication, data paper publication and science discovery publication. The three journals worked closely with authors to publish discovery papers as well as datasets and data papers. Finally, both SEDAC's and IGS's infrastructures have been supported by the National Aeronautics and Space Administration (NASA) EOSDIS community, and their extensive collections of knowledge and technical resources.

Geomagnetism data in Europe: the EPOS ERIC. Within the European geomagnetism community, the European Plate Observing System European Research Infrastructure Consortium (EPOS ERIC) has played a major role in promoting the uptake of 21st century technologies and standards to create more granular and robust metadata and dataset documentation (Chambodut et al. 2018; Flower and TGS Geomagnetic Observations 2019). Following EPOS ERIC's leadership, ISGI plans to migrate the repository's metadata records into an interoperable schema that will allow repositories to serve metadata to European aggregators like OpenAIRE. Currently, ISGI is considering implementing CERIF, DataCite, and/or DCAT compliant metadata. Since 2013, INTERMAGNET has been publishing yearly definitive data through the GFZ (GeoForschungsZentrum) Data Service, which serves dataset metadata to aggregators using various metadata standards and sharing protocols. Furthermore, a metadata development project is underway to gather metadata for all observatories recording geomagnetic data worldwide. This includes the INTERMAGNET geomagnetic observatories metadata combined with metadata records held by the WDC for Geomagnetism, Edinburgh.

The Chinese research data infrastructure. GCdataPR, NSSDC and WDC-RRE were among the original Chinese data repositories that joined the ICSU system of World Data Centers in 1988. In 2008, to promote collaboration between the eight Chinese repositories at the WDS, the WDS China Common Clearinghouse was created (Wang et al. 2020). The prototype for the WDS China's unified metadata search portal was constructed with Pycsw, a Python implementation of the OGC's Catalogue Services for the Web (CSW) specification (Wang et al. 2020). This initiative, led by WDC-RRE, encouraged and supported WDS members to develop harvestable metadata services based on similar spatial data interoperability standards, notably ISO 19115/19139/19119 metadata and the OGC CSW protocol.

Outside of the WDS, the Chinese repositories contribute to the larger Chinese digital research infrastructure, as part of 20 Chinese Data Centers organized under the National Science and Technology Infrastructure Center of China.[18] The 20 national data centers provide their metadata collections on a regular basis to a unified metadata search portal operated by the National Science and Technology Data Sharing Network of China (National Science and Technology Infrastructures 2016). These records must comply with the Chinese Science and Technology Infrastructure Resource Core Metadata standard (GB/T 30523-2014, China National Institute of Standardization 2014). Furthermore, all metadata records held by the 20 national data centers, including NSSDC, must be registered in accordance with the Science and Technology Resource Identification (CSTR), GB/T 32843-2016 (China National Institute of Standardization 2016), so that these metadata records can be discovered in the CSTR Identification platform. Another class of the data repository is peer reviewed dataset publications through the digital journal. The Global Change Data Repository is a digital journal (ISSN 2096-868X), which is issued monthly and compatible with the Journal of Global Change Data & Discovery (ISSN 2096-3645), a journal for publishing data papers. The two journals and the data and knowledge hub (metadata based links for specific applications) are part of the Global Change Research Data & Repository (GCdataPR). Through its publication methodology and procedures, the GCdataPR maintains long-term preservation and public availability of timely, quality and informative datasets. Both WDC-RRE and NSSDC also maintain custom metadata profiles that integrate local and international interoperability features. In addition, the China National Knowledge Infrastructure (CNKI) also is aggregating metadata from GCdataPR and WDC-RRE.

18 Many of which also maintain a close collaboration with WDS as non-members.

EOSDIS at NASA. Two of the repositories, SEDAC and IGS, are (at least partially) based in the United States, and they receive support from the National Aeronautics and Space Administration's (NASA) infrastructure. As one of NASA's Distributed Active Archive Centers (DAACs), SEDAC participates actively in initiatives stewarded by the Earth Science Data and Information System (ESDIS) project and SEDAC metadata is provided to NASA's EOSDIS Common Metadata Repository (CMR). The CMR is the back-end of Earthdata Search, the Global Change Master Directory (GCMD), and the International Data Network (IDN), the latter of which transfers SEDAC metadata into GEOSS. The complete collection of IGS data, which is distributed across data centers, has one of two complete mirrors hosted by a NASA EOSDIS data center, the Crustal Dynamics Data Information System (CDDIS) (the second mirror is hosted by the European Space Agency).[19] Thus, at present, metadata records for SEDAC datasets and for IGS collections are served in metadata search/retrieval endpoints at the CMR (Noll and Michael 2019), and they are available in multiple established metadata formats, specifically: DIF 10, ECHO 10, ISO 19115-2:2009 (MENDS and SMAP dialects), and UMM-C (Reiter and Ilincione 2019).

4. CHALLENGES

As described within the Methodology section, analyses and discussions of the similarities among the challenges that the HMetS-WG repositories face for developing and deploying harvestable metadata services led to consensus on the similarities observed among these challenges. The emerging consensus among the challenges that were reported by the repositories revealed three major overarching themes for the common challenges that were identified. The themes that represent the common challenges for developing and deploying harvestable metadata services include changing user needs, sustainability, and evolving technologies.

The three themes that were found for the challenges faced by the HMets-WG repositories when developing and deploying harvestable metadata services are closely linked to each other. Developing a good understanding of current and evolving technology trends and changing user needs, in light of existing and projected capabilities and resources, can help repositories to identify a sustainable approach for their new development efforts, and reduce the potential of incurring costs to employ expensive corrective measures in the future.

4.1. CHANGING USER NEEDS

The first major theme reflects repositories' efforts to identify and meet the changing needs of the user communities that they serve. Such efforts include adopting standards that maximize metadata interoperability, deploying metadata schemas that are widely used, but also versatile and extensible to address the changing needs of the user community. Serving the needs of repository users, including data producers and data reusers, is one of the primary objectives of research data repositories. Meeting the challenges for providing services to the user community as the needs of the users change is a key indicator of repository success.

Minimally, a research data repository exists to make a collection of data assets available to a designated community of users. Deploying harvestable metadata catalogues is a key strategy for reaching users, as these services can inform potential users and increase awareness of repository holdings. Such catalogues can be especially effective if they are tailored for interoperability with infrastructures (e.g. metacatalogues)[20] that are highly visible, feature-rich, widely-used, and also themselves integrated within the larger ecosystem of research infrastructures.

4.1.1. New users, new challenges

As a repository shares data more widely, its users become more diverse and heterogeneous. Catering to these evolving user needs is one of the most salient challenges faced by the HMetS-WG data repositories.

19 Since 2018, the European Space Agency (ESA) GNSS Science Support Centre (GSSC), located in Madrid started an initiative to create additional mirrors, in a project that also contemplates a new service platform: the IGS Global Data Centre (GNSS Science Support Centre 2023).

20 'A metacatalogue is a catalogue that allows for bibliographic searches in multiple catalogues.' See CRFCB 2014 and Kapiszewski and Karcher 2020.

ISGI and INTERMAGNET provide good examples of how users' growing diversity may pose challenges to repositories, even those with well-established data-sharing cultures. Open Data and sharing have always been essential for the geomagnetism community, as earth-observation research can rarely be done without data from multiple countries. In fact, geomagnetism's established data-sharing tradition is evidenced by over 50 years of collaborative data practices which have included yearly data publications and established, shared standards; e.g. the IAGA2002 data Exchange Format (2016). At INTERMAGNET, participants are volunteer magnetic observatories which, following standards defined by the network, seek to share and confidently reuse geomagnetic data within the community. ISGI's participants, in contrast, are institutes whose official task is defined by the International Association of Geomagnetism and Aeronomy (IAGA): to derive and make available officially endorsed data products. In recent years, the geomagnetism community has sought to achieve interoperability with other scientific fields of Earth and environmental observation, and to keep up with current trends to make data more usable, and also more useful, to a larger group of users, not only geomagnetism specialists. As we shall see below, both organizations have needed to factor in these developments when selecting their data and metadata sharing technologies.

Post-Pandemic, data driven regional economic development efforts involve new challenges, especially in rural areas, mountain regions, and small islands. In order to help such regional stakeholders, including decision makers and small business companies, GCdataPR initiated the Geographical Indications Environment & Sustainability (GIES) program. By opening quality datasets, data papers, and metadata (physical geographical data, agriculture products data, socio-economic data and local culture information, as well as in situ timely ecosystem monitoring data), the geographical indications or specific agriculture products could be used by consumers. The GIES cases clusters and practices demonstrated this as an effective solution for the repository to serve local people in attaining the 2030 Sustainable Development Goals (SDGs) (Liu, Gong & Liu, et al. 2021).

4.1.2. Stakeholder engagement, user outreach, adaptation of services

The repositories in this study have shown a clear user orientation, and most report an intent to serve diverse user communities: from the general public to industry data users, to researchers in highly specialized knowledge areas (see Table 2). Concerted outreach is regularly carried out among multiple groups of users and stakeholders, including current and potential users. Also, without exception, each of the repositories participates actively in sundry working groups and opportunities to exchange knowledge, within grassroots, top-down, or federated organizations. Some of these include the WDS, the Research Data Alliance (RDA), and the International Science Council's Committee on Data for Science and Technology (CODATA), the American Geophysical Union (AGU), the European Open Science Cloud's (EOSC) EPOS ERIC, the Group on Earth Observations (GEO), China-GEOSS, and the ESDIS system at NASA.

At IGS, for example, data services are being developed to meet the needs of new and established users (Ventura-Traveset, Navarro & Romero 2019), such as those found within IGS itself, including product coordinators, participants in working groups and pilot projects or in analysis centers, (Villiger & Dach 2019a: 139). But because all users of modern mapping, orientation and navigation systems are beneficiaries of the work done by IGS, the IGS Central Bureau has established various channels for outreach and communication (ibid.: 18), with the public and individuals, enterprises, non-profits, institutions and government actors worldwide. These channels include social media outlets like Twitter, where IGS uses the #GNSS4impact hashtag to tweet about common applications of GNSS data. Part of the aim is to make the general public aware of this foundational yet invisible infrastructure. Making IGSs work visible to the general public in ways that can be measured – such as through citation of IGS data, products, and other published outputs – helps IGS advocate for the organization and make a strong case to its supporting partners and funders (IGS Central Bureau 2023).

Repositories also will need to adapt the metadata that they distribute to address the current needs of the user communities that they serve as these needs change. In addition to revising repository services offered, such as recommended uses and data formats and the like, it may be necessary to adopt metadata standards and enhance metadata harvesting capabilities to reflect the knowledge and research interests of the new community segments and domains that are being served. For example, a repository may discover changes in the disciplines of its users by

identifying the disciplines of publications and authors that are currently citing the repository's data holdings. Learning about such changes can enable the repository to identify additional metadata standards, particular metadata elements, specific vocabularies and harvesters that can serve the needs of the new communities as the disciplines of users change. Recent developments, such as those described by Musen et. al (2022), include metadata templates, discipline-specific ontologies, and metadata evaluation software tools that enable rich FAIR-compliant metadata to be produced for distribution to particular communities and across communities of data users.

4.1.3. Repository usage metrics and citation counts

To some extent, repositories can keep track of their efforts to increase data discovery and, ultimately usage, through counters that measure user engagement with repository assets (e.g. clicks, downloads, searches, turnaways), which can help keep track of fluctuations and patterns in a repository's engagement and usage. A current standard for repository metrics is embodied in the COUNTER Code of Practice (Fenner et al. 2018). Some repositories, such as NSSDC and SEDAC employ a simple user authentication requirement, via a single log-in or registration with an e-mail address, to gain insight into data usage patterns beyond raw metrics, shedding light onto the frequency of usage for each item and the types of users who may be accessing data assets. In contrast, GCdataPR reports using IP addresses and real-time usage statistics to keep track of the repository's international visits, in a way that is consistent with GCdataPR's stated goal of reaching a broader international user base. But, while potentially useful for tracking users' online interactions with the repository, these alternative metrics also have limitations as indicators of actual dataset reuse (Ramachandran, Bugbee and Murphy, 2021).

Alternatively, data citation tracking, despite its limitations,[21] is increasingly becoming a tool that can be used to estimate the scientific impact of a repository's data assets and to facilitate some types of bibliometric analysis of data usage. Among our use cases, GCdataPR, SEDAC and WDC-RRE report tracking data citations. SEDACs platform has also implemented a searchable online database that contains references to citations of the repository's datasets (Socioeconomic Data and Applications Center 2023a).

4.2. SUSTAINABILITY

The second set of challenges of repositories for developing and deploying harvestable metadata services refers to the ways in which repositories are limited in terms of opportunities for ensuring the sustainability of their services, especially when considering resource and policy constraints. Sustainable services are needed to provide continuous operations while facing the combined challenges of meeting the changing needs of users with technology that is evolving. Furthermore, with limited resources for technical development, repositories must consider the costs of establishing new services while providing and maintaining existing services.

Securing continual support for sustainable repository development and maintenance is a fundamental management challenge, especially for small- and medium-scale research facilities. Our group of repositories have faced these challenges by gaining support within their host institutions and finding support through partnerships.

4.2.1. Sustainable growth and operations

In research organizations without a strong culture of research data management (RDM), it may take time to build support for expanding data services with initiatives such as a new metadata service. For example, the Göittingen eResearch Alliance (Dierkes & Wuttke 2016) built institutional support by engaging with the organization's key decision makers and stakeholders. Alternatively, SEDAC and the three WDS members in China have been able to build support for their data centers within their host institutions and their national data infrastructures, and this is reflected in the repositories' maturity status. These examples also underscore that collaboration among community stakeholders fosters efforts to attain data repository interoperability, as reported by Gries et al. (2018).

21 Citing data is still a relatively new practice, not yet practiced consistently across the research community. For example, one recent study of 12 324 COVID-19-related articles (Zuo et al. 2021) reports that 28.5% provided at least one URL for a dataset that had been (re)used in the article. Although the article does not quantify the difference, it reports that data citation formats were also heterogeneous: some authors provided in-text URLs only, while others gave full bibliographic references.

For less hierarchical organizations like research networks and data federations, the most salient challenges involve coordinating the development of a common standard or application profile, or coordinating the adoption of an existing technology (Yarmey & Baker 2013). The two WDS network members among our use cases, IGS and INTERMAGNET, are different examples of established, international data federations that managed to create impressive infrastructures on the basis of voluntary member participation, through many decades of collaborative work.

The voluntary, federated character of IGS relies on decentralized funding schemes for projects and initiatives, usually by public institutions, governments or other research organizations. To maintain its reliable service provision, IGS must rely on system redundancy and on multi-year support commitments from the institutions that host the key elements of the system (IGS Central Bureau 2023; Villiger & Dach 2019b). To marshal support for a project, repository partners have to be able to envision the positive and tangible ways in which the project will impact funding partners and their constituencies, and how it will benefit the institution and society as a whole. In particular, IGS public outreach and communication initiatives reflect the organizations keen understanding of that fact.

4.2.2. Resource constraints

It is also useful to bear in mind that open-source software (OSS) is being produced and made available on a regular basis, some of which is intended for repositories to implement harvesting protocols with lower investment costs. For example, harvesting protocols can be implemented as modules in bespoke repository platforms by means of Viringo, an OAI-PMH API created by DataCite and further developed at FRDR, or Pycsw, a Python implementation of the OGC CSW protocol that is used by WDCRRE for its Catalogue Service. A minimal implementation of harvestable metadata may consist of a web-accessible folder (WAF), sitemap, or publicly accessible XML file of machine-readable metadata.

While a discussion of the advantages and disadvantages of OSS lies outside of the scope of this paper (see Trappler 2009, for a discussion of OSS pros and cons), it bears mentioning that repository managers will need to weigh the benefits of OSS against potential trade-offs (e.g. increased labor costs, community vs. corporate support services, etc.). Nevertheless, software solutions implemented with OSS may offer advantages for adoption if technological compatibility and software reusability is possible.

Independent of the decision to select a particular approach for implementing an enhancement, such as harvestable metadata capabilities, additional sources of support may be needed to sustainably develop and deploy improvements to data repository infrastructure. If the costs of enhancements are not absorbed by operating budgets, such costs may need to be supported separately. In such cases, data repositories may need to initiate projects and secure additional support for improvements to their services as part of their approach to providing sustainable data stewardship (Downs & Chen, 2016).

4.3. EVOLVING TECHNOLOGIES

The third theme reflects the set of challenges for making strategic decisions and associated investments in a landscape of evolving technologies and changing standards. Weighing the factors that influence such decisions presents a significant challenge for repository managers. Repositories must assess the potential of a technology or standard to meet current and future needs, as well as its maturity, to determine whether and when it can be adopted.

The repositories in this study represent established data-sharing communities that have been sharing scientific data (in analogue and digital formats) long before the advent of the internet. Considering the ever-changing technological landscape, the 'ideal' constellation of technologies and services may seem like a moving target: Over the past few decades, these repositories have experienced multiple waves of technical innovation, which have time and again transformed the ways in which data is obtained, documented and shared with other researchers.

4.3.1. Metadata and open data access policies

In general, repositories may be hesitant to expose metadata for protected datasets and/ or collections. Although none of our repositories reported hosting private or confidential data, some assets in the NSSDC repository are embargoed for a short time period, which is deemed long enough to ensure that data owners' rights and interests are protected. NSSDC's approach

is compatible with the requirement that data be as open as allowable, but as restricted as necessary. SEDAC favors the use of open data licenses (mainly CC BY 4.0),[22] 'unless there are extenuating circumstances such as data restrictions inherited from input data' (Socioeconomic Data and Applications Center 2023b). Wherever relevant, necessary consideration must also be given to data sharing practices and principles – beyond FAIR – that focus on various ethical concerns, such as the First Nations Principles of OCAP (First Nations Information Governance Centre 2014), and the CARE Principles for Indigenous Data Governance (Carroll et al. 2020). This means investing in the technical solutions that embody those principles: differentiated access policies and secure data storage, with trustworthy capabilities for offering selective data access under distinct protection classifications; or providing access only to authorized users. Machine-readable data licenses in metadata (Creative Commons 2002) can instruct search engines and automated software to display and filter content according to their licensing, which can in turn remind users of the freedoms and obligations (e.g. proper attribution) associated with the dataset.

4.3.2. PIDs, DOIs, and identifiers for dynamic datasets

Persistent, unique identifiers (PIDs) for digital objects can enhance and enable a range of interoperability features, from automatic metadata retrieval for bibliographic references in tools like RefWorks and Zotero, to deduplicated aggregation of dataset metadata into federated catalogues, to the analysis and visualization of networks of scholarly communication and collaboration like OpenAIRE's Research Graph (Manghi et al. 2019). The Digital Object Identifier (DOI) standard (Paskin 1999), which emerged in the 1990s, as well as newer PIDs like the Research Organization Registry (ROR)[23] and Open Researcher and Contributor IDs (ORCID)[24] have opened new avenues for automating links between metadata records, and for creating new digital research services. The growing use of the ROR identifier in dataset metadata is a case in point. Since implementing ROR tags in 2020, national aggregation platforms like the Federated Research Data Repository (FRDR) have the option to selectively harvest Canadian data from non-Canadian repositories when at least one of the authors is affiliated with a Canadian research organization (Digital Research Alliance 2023). Similarly, ORCIDs make it easier to track the scholarly output of individual researchers.

The ability to permanently and uniquely reference arbitrary data subsets and subsequent versions of a dataset is key to safeguarding the reproducibility of scientific studies that rely on shared data. To tackle the technical challenge involved, groups such as DataCite (DataCite Metadata Working Group 2021) and the Research Data Alliance's (RDA) Data Versioning Working Group (Klump et al. 2021; Klump et al. 2020) have developed approaches and recommendations to implement dataset versioning and dynamic data citation. In 2015 the latter group released an RDA recommendation describing the dynamic assignment of PIDs to every new, unique data query that produced a given data subset (Rauber et al. 2015). With this approach, when a dataset changes due to updates or reprocessing (Klump et al. 2021), or when a subset of data is extracted from a larger dataset, or republished within a larger data collection (as described in Klump, Huber & Diepenbroek 2016), these unique products can themselves be reconstructed identified, referenced, cited and reused. These RDA recommendations have been implemented in various data repositories that enable citation of time-stamped versions of subsetted dynamic datasets with persistent identifiers, facilitating retrieval, across sundry data types, for reuse (Rauber et al. 2021).

Of our present set of use cases, only the WDC-RRE repository reported having already implemented a system to assign PIDs to versioned datasets (WDC-RRE 2016), in which identifiers are coded to refer back to data queries executed on specific, timestamped dataset versions. Two others, INTERMAGNET and ISGI, expressed an interest in developing a PID versioning system in future stages of their repositories' development. This approach would expedite the release of non-definitive datasets of geomagnetic observations, making these very detailed and highly valuable data assets available sooner to the scientific community. Another recent and well-documented example of metadata versioning from a WDS Member repository is Project MINTED at Ocean Networks Canada (Jenkyns & Ridsdale 2020; Jenkyns 2019), who also had an active role in developing the RDA's Data Versioning WG's outputs.

22 Creative Commons Attribution 4.0 International, https://creativecommons.org/licenses/by/4.0/.

23 Research Organization Registry, https://ror.org/about/.

24 Open Researcher and Contributor ID, https://info.orcid.org/what-is-orcid/.

4.3.3. Maximizing data asset potential: Two approaches

To determine how much an existing repository infrastructure can achieve, and to pursue new development opportunities accordingly, an ongoing and thorough assessment of a repository's infrastructure is recommended. To support a repository's initial self-assessment, the global RDM community has produced instruments to assess the maturity and trustworthiness of a data repository and the data assets, including metadata records, it contains (Downs 2021; Peng 2018). Some practical, up-to-date frameworks for reviewing a repository's current state are the most recent version of the CoreTrustSeal requirements for trustworthy data repositories (CoreTrustSeal 2022), the RDAs new FAIR data maturity model (FAIR Data Maturity Model WG 2020), the CARE Principles for Indigenous Data Governance (Carroll et al. 2020), and the TRUST Principles for digital repositories (Lin et al. 2020). Data repositories also need to continually assess the technology landscape to identify opportunities for improving capabilities to serve their designated communities. Cooperating with other repositories, within and across disciplines, helps with such assessments, especially when cooperating repositories share adoption stories and lessons-learned.

Two cases in our study reflect an interplay between changing user needs, evolving technologies, and resource constraints. The two geomagnetism data repositories, INTERMAGNET and ISGI, contain data assets with enormous potential for innovative, interdisciplinary research, but whose metadata formats and services have not been updated to current standards. For each repository, the challenge lies in finding a strategy that will allow them to exploit their data's potential to serve their current (known) users as well as future (known and unknown) ones. It involves optimizing between general and use-case based repository developments, including metadata standards and exchange protocols. ISGI and INTERMAGNET have reported different strategies, based on different priorities, to respond to this challenge. INTERMAGNET has reported having to ponder the advantages of general-purpose, extensive standards that can open future (yet unknown) avenues of research and collaboration, versus use-case based approaches that tailor new developments to better support each new case. In contrast, the existence of concrete opportunities for interdisciplinary collaboration for example, between ISGI and researchers in the biological sciences may justify an approach that tailors a repository's developments to a set of concrete use cases, taking a chance on their potential for future extensibility. HMetS-WG repositories also recognize the tension between the two fundamental principles of investing in future-proof technologies or maximizing user engagement with the data over time. In practice, repositories will usually attempt to balance both principles when designing their development plans.

4.4. LIMITATIONS OF A HARVESTING STRATEGY FOR DATASET DISCOVERY

In many of the cases described in these reports, the development strategy for harvestable metadata services has been very thorough. To varying degrees, the SEDAC, WDC-RRE and GCdataPR use cases hint at the limits of a discovery/findability strategy based on harvestable metadata services alone. These repositories, in particular, have motivated the ITOs decision to create an inventory of metadata aggregation services (Li & Payne 2021) that will allow repository managers to find aggregators outside their community's beaten path. Furthermore, and as mentioned above, motivated in part by inclusion in Google Dataset Search, SEDAC has a metadata harvesting capability already underway. Furthermore, WDC-RRE and GCdataPR have expressed future interest in receiving ITO support to develop a semantic metadata strategy as well.

5. CONCLUSIONS

The experiences reported in this study frame the socio-technical dimensions of research service development, where success depends largely on meeting the diverse needs of stakeholders within the designated communities of the repositories studied. And within each repository, the users may reflect different research perspectives in terms of interests and methods, or they may even employ different epistemological and ontological approaches (Poirier & Costelloe-Kuehn 2019). In effect, developing repository services, including harvestable metadata, involves identifying, adopting, and developing technologies that are continuously evolving to demonstrably serve the changing needs of heterogeneous user communities, within the policy and funding constraints of the institution. While the 'ideal' constellation of technologies and services may seem like a moving target, finding the right balance for their unique use case appears to be an attainable goal for most repositories.

When developing new services using cross-domain recommendations and policies, the "need for standardization and interoperability" must be balanced 'against the need for flexibility and discipline-specific nuance' (Goddard et al. 2021). Which standards and technologies will best serve the original producers and established users of datasets, as well as the larger user community, including new and future data users? Nearly all of our repositories conduct some level of market research and intelligence gathering to inform their service development in general, and harvestable metadata services in particular: Gathering usage data and data citation counts and characteristics is necessary to monitor how data is queried and used. Other common practices involve engaging in designated community outreach and participation in cross-domain and/or international working groups, as well as having dedicated working groups with diverse stakeholders; or engaging with current and prospective users directly, such as via interdisciplinary research collaborations.

Strategies for project sustainability vary according to the repositories' institutional structure. For repositories embedded in centralized and hierarchical institutions (such as research centers, or national digital infrastructure projects), attaining long-term sustainability is contingent on continued support by parent organizations. In these settings, some key strategies include sustained engagement with the organization's key decision makers and stakeholders to seek strategic alignment, and maximizing opportunities to build support for data centers within their host institutions. For repositories embedded in decentralized organizations, like research networks and data federations, the main sustainability challenge is one of coordination and community development. Among our use cases, IGS and INTERMAGNET represent examples of data infrastructures that leverage voluntary member participation and decades of collaborative work to develop and maintain their services over time.

Lastly, the results from this study strongly suggest that participation and integration into technical networks (national, regional or subject-specific) can be a driver of technological development in member repositories. In all cases, the intermediating entity (a network, community or institution) effectively functions as a catalyst for service development and standards implementation, as well as an incubator that connects repositories' local ecosystems with global research data sharing spaces. The three themes that have been identified in this study for the challenges of developing and deploying harvestable metadata services also offer implications for the challenges that repositories face, generally and in terms of other capabilities, as they try to improve their services while meeting the changing needs of users with evolving technology in a sustainable manner. Such implications may be considerations for future research and theory development.

ACKNOWLEDGEMENTS

The authors appreciate contributions to the use cases received from Robert S. Chen and Sri Vinay of SEDAC; Mayra I. Oyola of IGS; Bu Kun of WDC-RRE; Sarah Reay of INTERMAGNET and the WDS-Geomagnetism (Edinburgh); Ruixiang Shi, Junhua Ma, Yinghua Zhang of GCdataPR; and Winnie Li of WDS-ITO. The thoughtful comments and edits that Sarah Reay also contributed to this manuscript are very much appreciated as well. For their guest presentations in the working group, our thanks go to Alex Garnett from the Federated Research Data Repository in Canada (FRDR); JJ Kavelaars from the Canadian Astronomy Data Centre and the International Virtual Observatory Alliance (IVOA); Daniella Lowenberg from the California Digital Library and Project COUNTER/Make Data Count; and Paolo Manghi from OpenAIRE. We'd like to acknowledge the collegial support by the Portage Network and members of the Discovery & Metadata Expert Group (DMEG). Many thanks especially to Kelly Stathis, Kevin Read, Peter Webster, Lee Wilson, Amber Leahey, Eugene Barsky, and Jennifer Abel for their comments and feedback during various stages of the HMetS-WG project. The editorial assistance of Caroline Lee of the WDS-ITO also is appreciated. The authors appreciate the thoughtful recommendations for improvement of the manuscript that were offered by anonymous Data Science Journal reviewers.

DISCLOSURE STATEMENT

This work represents the perspectives of the authors and does not necessarily reflect those of their sponsors or employers.

FUNDING INFORMATION

Support for the Harvestable Metadata Services Group was funded by the Digital Research Alliance of Canada (formerly New Digital Research Infrastructure Organization, NDRIO) under Collaborative agreement number 51387-5032. A portion of the contributions of Alicia Urquidi Diaz was made possible by Scholars Portal of the Ontario Council of University Libraries (OCUL). The contributions of Robert R. Downs were supported by the National Aeronautics and Space Administration under Contract 80GSFC18C0111 for the Socioeconomic Data and Applications Distributed Active Archive Center (DAAC). The paper's publication was supported by the Global Change Data Publishing and Repository (GCdataPR).

COMPETING INTERESTS

Authors RRD, QX, JW, AC, LC, and SF represented the repositories that were the focus of the study while the study was being conducted. RRD is a DSJ Editorial Board Member.

AUTHOR CONTRIBUTIONS

AUD prepared the initial draft of the manuscript and all authors reviewed and revised the manuscript prior to submission. AUD, KP, RRD, and AC contributed to conceptualizing the study. AUD, RRD and KP designed and developed the methodology. All authors contributed to the research and investigation process. AUD and RRD revised the manuscript to address recommendations offered by the anonymous Data Science Journal reviewers and all authors reviewed the revised manuscript.

AUTHOR AFFILIATIONS

Robert R. Downs orcid.org/0000-0002-8595-5134
Center for International Earth Science Information Network (CIESIN), Columbia Climate School, Columbia University, United States

Alicia Urquidi Díaz orcid.org/0000-0001-9766-3545
World Data System – International Technology Office, Canada; Scholars Portal, Canada

Qi Xu orcid.org/0009-0006-3155-2475
National Space Science Data Center, China; National Space Science Data Center, Chinese Academy of Sciences, China

Juanle Wang orcid.org/0000-0002-5641-0813
WDC-Renewable Resources and Environment, China; Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, China

Aude Chambodut orcid.org/0000-0001-8793-1315
International Service of Geomagnetic Indices, France

Chuang Liu orcid.org/0000-0002-4728-670X
Global Change Research Data Publishing and Repository, China

Simon Flower orcid.org/0000-0002-8173-0181
WDC-Geomagnetism (Edinburgh), UK; INTERMAGNET, UK

Karen Payne orcid.org/0000-0003-0608-5378
Center for International Earth Science Information Network (CIESIN), Columbia Climate School, Columbia University, United States

REFERENCES

Bornatici, C, Kleiner, B, Kvamme, T, Kvalheim, V, Bradić-Martinović, A and **Glavica, M.** 2017. CESSDA SaW D3.3: Guide for national planning for setting up new data services (V4.0). *Zenodo*. [Last accessed 10 February 2023]. DOI: https://doi.org/10.5334/dsj-2020-043

Carroll, SR, Garba, I, Figueroa-Rodríguez, OL, Holbrook, J, Lovett, R, Materechera, S, Parsons, M, Raseroka, K, Rodriguez-Lonebear, D, Rowe, R, Sara, R, Walker, JD, Anderson, J and **Hudson, M.** 2020. The CARE Principles for Indigenous Data Governance. *Data Science Journal*, 19(1): 43. [Last accessed 1 February 2023]. DOI: http://doi.org/10.5334/dsj-2020-043

Chambodut, A, Flower, S, Hejda, P, Smirnov, M, Thomson, A and **Viljanon, A.** 2018. The Geomagnetic field between Earths core and space: how the Geomagnetic Observations 'EPOS Thematic Core Services (TCS) address the various sources of the Geomagnetic field'. In: Filosa, S (ed.) *The EPOS*

Newsletter 2. Roma Lazio, Italy: EPOS. Available at https://epos-eu.org/node/879/pdf [Last accessed 1 February 2023].

China National Institute of Standardization. 2014. GB/T 30523-2014 Science and technology infrastructure. *Resource Core Metadata*. National Standard Announcement 2014 No. 2. Available at https://www.chinesestandard.net/PDF/English.aspx/GBT30523-2014 [Last accessed 10 March 2023].

China National Institute of Standardization. 2014. GB/T 32843-2016. Science and Technology Resource Identification. National Standard Announcement 2016 No.14. Available at https://www.chinesestandard.net/PDF/English.aspx/GBT32843-2016 [Last accessed 10 March 2023].

CoreTrustSeal Standards and Certification Board. 2022. CoreTrustSeal Requirements 2023–2025 (V01.00). *Zenodo*. [Last accessed 1 February 2023]. DOI: https://doi.org/10.5281/zenodo.7051012

Creative Commons. 2002. Creative Commons Unveils Machine-Readable Licenses. Portal. Available at https://creativecommons.org/2002/12/16/creativecommonsunveilsmachinereadablecopyrightlicenses/ [Last accessed 1 February 2023].

CRFCB. Jan. 2014. Metacatalogue. Available at https://blogs.univ-poitiers.fr/glossaire-mco/2014/01/23/metacatalogue/ [Last accessed 1 February 2023].

Culina, A, Baglioni, M, Crowther, TW, Visser, ME, Woutersen-Windhouwer, S and **Manghi, P.** 2018. Navigating the unfolding open data landscape in ecology and evolution. *Nature Ecology & Evolution*, 2(3): 420–426. [Last accessed 1 February 2023]. DOI: https://doi.org/10.1038/s41559-017-0458-2

DataCite Metadata Working Group. 2021. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.4. DataCite e.V. [Last accessed 1 February 2023]. DOI: https://doi.org/10.14454/3w3z-sa82

DCMI. 2020. Dublin Core Metadata Initiative. Available at https://www.dublincore.org/ [Last accessed 15 October 2023].

Dierkes, J and **Wuttke, U.** 2016. The Göttingen eResearch Alliance: A Case Study of Developing and Establishing Institutional Support for Research Data Management. *ISPRS International Journal of Geo-Information*, 5(8): 133. [Last accessed 1 February 2023]. DOI: https://doi.org/10.3390/ijgi5080133

Dietze, MC, Fox, A, Beck-Johnson, LM and **White, EP.** 2018. Iterative near-term ecological forecasting: Needs, opportunities, and challenges. *Proceedings of the National Academy of Sciences* 115.7, 1424–1432. [Last accessed 1 February 2023]. DOI: https://doi.org/10.1073/pnas.1710231115

Digital Research Alliance. 2023. Research Data Management. Available at https://alliancecan.ca/en/services/research-data-management [Last accessed 10 February 2023].

Downs, RR. 2021. Improving opportunities for new value of open data: Assessing and certifying research data repositories. *Data Science Journal*, 20(1): 1. [Last accessed 1 February 2023]. DOI: https://doi.org/10.5334/dsj-2021-001

Downs, RR and **Chen, RS.** 2016. A Portfolio Approach to a Sustainable Business Model for Scientific Data Stewardship. *SciDataCon 2016 Conference Paper*. Denver, Colorado, 11–13 September 2016. DOI: https://doi.org/10.7916/d8-fae5-cz67

FAIR Data Maturity Model Working Group. 2020. FAIR Data Maturity Model. Specification and Guidelines (1.0). [Last accessed 1 February 2023]. DOI: https://doi.org/10.15497/rda00050

Fenner, M, Lowenberg, D, Jones, M, Needham, P, Vieglas, D, Abrams, S, Cruse, P and **Chodaki, J.** 2018. Code of practice for research data usage. Metrics release 1 (pp. 1–38). Make Date Count Project. [Accessed 1 February 2023]. DOI: https://doi.org/10.5281/zenodo.3340591

Flower, S and **TGS Geomagnetic Observations.** 2019. EPOS and Geomagnetism, together for the benefit of the community! In: *The EPOS Newsletter Special Issue 04*. Available at https://mailchi.mp/a62aff0adf0c/specialissue-n-04-2019-3057861 [Last accessed 1 February 2023].

First Nations Information Governance Centre. 2014. Barriers and levers for the implementation of OCAP. *International Indigenous Policy Journal*, 2(3). [Last accessed 1 February 2023]. DOI: https://doi.org/10/gkcx7m

Garnett, A, Leahey, A, Savard, D, Towell, B and **Wilson, L.** 2017. Open metadata for research data discovery in Canada. *Journal of Library Metadata*, 17(3–4): 201–217. [Last accessed 1 February 2023]. DOI: https://doi.org/10.1080/19386389.2018.1443698

Goddard, L, Khair, S, Higgins, S and **Doiron, J.** 2021. RDM for Digitally-Curious Humanists. *OSF*. [Last accessed 1 February 2023]. DOI: https://doi.org/10.17605/OSF.IO/6VEPJ

GNSS Science Support Centre. 2023. European Space Agency. https://gssc.esa.int/portal/datasets [Last accessed 1 February 2023].

Gries, C, Budden, A, Laney, C, O'Brien, M, Servilla, M, Sheldon, W, Vanderbilt, K and **Vieglais, D.** 2018. Facilitating and improving environmental research data repository interoperability. *Data Science Journal*, 17: 22. [Last accessed 6 February 2023]. DOI: http://doi.org/10.5334/dsj-2018-022

IAGA2002 Exchange Format. 2016. NOAA National Centers for Environmental Information (NCEI). Available at https://www.ngdc.noaa.gov/IAGA/vdat/IAGA2002/iaga2002format.html [Last accessed 10 March 2023].

IGS Central Bureau. 2023. Data. Available at https://igs.org/data/ [Last accessed 10 March 2023].

International Standards Office. 2019. ISO 19115–1:2014. Geographic information — Metadata — Part 1: Fundamentals. Available at https://www.iso.org/standard/53798.html [Last accessed 1 February 2023].

Jenkyns, R. 2019. Making Identifiers Necessary to Track Evolving Data (MINTED) – A Brief Overview. National Data Services Framework Summit 2019 (NDSF 2019), Ottawa, Canada. *Zenodo*. [Accessed 1 February 2023]. DOI: https://doi.org/10.5281/zenodo.2555355

Jenkyns, R and **Ridsdale, C.** 2020. MINTED: Making Identifiers Necessary for Tracking Evolving Data. CANARIE Research Data Management Workshop 2020, Kanata, Canada. *Zenodo*. [Last accessed 1 February 2023]. DOI: https://doi.org/10.5281/zenodo.3665755

JISC. 2011. JISC Project Plan Template. p. 15. Available at https://www.slideshare.net/butest/jisc-project-plan-template-3859125 or https://www.yumpu.com/en/document/view/23470278/jisc-project-plan-template [Last accessed 8 February 2023].

Kapiszewski, D and **Karcher. S.** 2020. Making Research Data Accessible. In: Elman, C, Mahoney, J, and Gerring J (eds.), *The Production of Knowledge: Enhancing Progress in Social Science*, Strategies for Social Inquiry. Cambridge: Cambridge University Press, pp. 197–220. [Last accessed 1 February 2023]. DOI: https://doi.org/10.1017/9781108762519.008

Klump, J, Huber, R and **Diepenbroek, M.** 2016. DOI for geoscience data – how early practices shape present perceptions. *Earth Science Informatics*, 9: 123–136. [Last accessed 5 October 2021]. DOI: https://doi.org/10.1007/s12145-015-0231-5

Klump, J, Wyborn, L, Downs, R, Asmi, A, Wu, M, Ryder, G and **Martin, J.** 2020. Data Versioning Working Group Updated Compilation of Data Versioning Use Cases (1.1). [Last accessed 1 February 2023]. DOI: https://doi.org/10.15497/RDA00041

Klump, J, Wyborn, L, Wu, M, Martin, J, Downs, RR and **Asmi, A.** 2021. Versioning data is about more than revisions: A conceptual framework and proposed principles. *Data Science Journal*, 20(1): 12. [Last accessed 1 February 2023]. DOI: https://doi.org/10.5334/dsj-2021-012

Kramer, T, Klas, C and **Hausstein, B.** 2018. A data discovery index for the social sciences. *Scientific Data*, 5: 1–10. [Last accessed 1 February 2023]. DOI: https://doi.org/10.1038/sdata.2018.64

Lagoze, C, Van de Sompel, H, Nelson, M and **Warner, S.** 2005. Implementation Guidelines for the Open Archives Initiative Protocol for Metadata Harvesting: Guidelines for Repository Implementers. Available at https://www.openarchives.org/OAI/2.0/guidelines-repository.htm [Last accessed 10 February 2023].

Li, WAW and **Payne, K.** 2021. Searchable Index of Metadata Aggregators [Data set]. *Zenodo*. [Last accessed 1 February 2023]. DOI: https://doi.org/10.5281/zenodo.4589050

Lin, D, Crabtree, J, Dillo, I, Downs, RR, Edmunds, R, Giaretta, D, De Giusiti, M, L'Hours, H, Hugo, W, Jenkyns, R, Khodiyar, V, Martone, M, Mokrane, M, Navale, V, Petters, J, Sierman, B, Sokolova, DV, Stockhause, M and **Westbrook, J.** 2020. The TRUST Principles for Digital Repositories. *Scientific Data*, 7: 144. [Last accessed 1 February 2023]. DOI: https://doi.org/10.1038/s41597-020-0486-7

Liu, C, Gong, K, Liu, YH , Liao, XH, Wang, ZB, He, CC, Luo, H, Zhou, X, Tong, QX, Min, QW, Wu, JJ, Gui, DW, Chen, J, Wang, PP, Lu, F, Zhou, L, Sun, YW, Yang, X, Li, J, Wang, XQ, Tian, H, Zhang, GY, Chen, CX, Guo, P, Liang, Y, Xu, GC, Zhang, ZX, Yu, XY, Zhang, XD, Issa, AM, Song, XF, Wang, ZX, Fu, JY, Wang, YS, Zhu, XG, Zhang, LF, Zhu, YQ, Yu, BH, Wang, G, Lin, G, Dai, X and **Lyv, YH.** 2021. An innovative solution on geographical indications for environment & sustainability (GIES) [J]. *Journal of Global Change Data & Discovery*, 5(3): 237–248. [Last accessed 10 March 2023]. DOI: https://doi.org/10.3974/geodp.2021.03.03

Lokers, R, Knapen, R, Janssen, S, van Randen, Y and **Jansen, J.** 2016. Analysis of Big Data technologies for use in agro-environmental science. *Environmental Modelling & Software*, 84: 494–504. [Last accessed 1 February 2023]. DOI: https://doi.org/10.1016/j.envsoft.2016.07.017

Manghi, P, Bardi, A, Atzori, C, Baglioni, M, Manola, N, Schirrwagen, J and **Principe, P.** 2019. The OpenAIRE Research Graph Data Model (1.3). *Zenodo*. [Last accessed 1 February 2023]. DOI: https://doi.org/10.5281/zenodo.2643199

Musen, MA, O'Connor, MJ, Schultes, E, Martínez-Romero, M, Hardi, J and **Graybeal, J.** 2022. Modeling community standards for metadata as templates makes data FAIR. *Scientific Data*, 9(1): 696. [Last accessed 7 February 2023]. DOI: https://doi.org/10.1038/s41597-022-01815-3

National Science and Technology Infrastructures. 2016. International Research Collaboration Information Platform. Available at http://www.ircip.cn/web/999722-999726.html?id=26645&newsid=645574 [Last accessed 8 February 2023].

Nebert, D, Voges, U, Bigagli, L, Panagiotis, V and **Westcott, B.** 2016. OGC Catalogue Services 3.0 – General Model. Available at https://docs.opengeospatial.org/is/12-168r6/12-168r6.html [Last accessed 1 February 2023].

Noll, C and **Michael, P.** 2019. CDDIS Global Data Center Technical Report 2019. NASA Goddard Space Flight Center. In: *2019 IGS Technical Report*. https://cddis.nasa.gov/docs/2019/CDDIStr_Noll_2019_v3.pdf [Last accessed 1 February 2023].

Paskin, N. 1999. Toward unique identifiers. *Proceedings of the IEEE*, 87(7): 1208–1227. [Last accessed 1 February 2023]. DOI: https://doi.org/10.1109/5.771073

Payne, K and **Urquidi Diaz, A.** 2020. World Data System Member Survey 2019. *Zenodo*. [Last accessed 1 February 2023]. DOI: https://doi.org/10.5281/zenodo.3840406

Payne, K, Urquidi Diaz, A and **Li, WAW.** 2021. Steps towards an harvestable metadata services implementation plan (flow-chart format) & template: Metadata service implementation plan (1.1). *Zenodo*. [Last accessed 8 February 2023]. DOI: https://doi.org/10.5281/zenodo.4589013

Payne, K and **Verhey, C.** 2022. How to use research data alliance crosswalks to recommend Schema.org markup in metadata. *Zenodo*. [Last accessed 1 February 2023]. DOI: https://doi.org/10.5281/zenodo.4589090

Peng, G. 2018. The state of assessing data stewardship maturity – An overview. *Data Science Journal*, 17: 7. [Last accessed 1 February 2023]. DOI: http://doi.org/10.5334/dsj-2018-007

Plante, RL, Becker, CA, Medina-Smith, A, Brady, K, Dima, A, Long, B, Bartolo, LM, Warren, JA and **Hanisch, RJ.** 2021. Implementing a registry federation for materials science data discovery. *Data Science Journal*, 20(1): 15. [Last accessed 1 February 2023]. DOI: http://doi.org/10.5334/dsj-2021-015

Poirier, L and **Costelloe-Kuehn, B.** 2019. Data sharing at scale: A heuristic for affirming data cultures. *Data Science Journal*, 18(1): 48. [Last accessed 1 February 2023]. DOI: http://doi.org/10.5334/dsj-2019-048

Ramachandran, R, Bugbee, K and **Murphy, K.** 2021. From open data to open science. *Earth and Space Science*, 8(5): e2020EA001562. [Last accessed 6 February 2023]. DOI: https://doi.org/10.1029/2020EA001562

Rauber, A, Asmi, A, van Uytvanck, D and **Proell, S.** 2015. Data citation of evolving data: Recommendations of the Working Group on Data Citation (WGDC). *Zenodo*. [Last accessed 1 February 2023]. DOI: https://doi.org/10.15497/RDA00016

Rauber, A, Gößwein, B, Zwölf, CM, Schubert, C, Wörister, F, Duncan, J, Flicker, K, Zettsu, K, Meixner, K, McIntosh, LD and **Jenkyns, R.** 2021. Precisely and persistently identifying and citing arbitrary subsets of dynamic data. *Harvard Data Science Review*, 3(4). [Last accessed 8 February 2023]. DOI: https://doi.org/10.1162/99608f92.be565013

Reiter, E and **Rindone, J.** 2019. CMR Data Partner User Guide. Available at https://wiki.earthdata.nasa.gov/display/CMR/CMR+Data+Partner+User+Guide#CMRDataPartnerUserGuide-DataPartnerTasks [Last accessed 1 February 2023].

Socioeconomic Data and Applications Center (SEDAC). 2023a Citations Database. Available at https://sedac.ciesin.columbia.edu/citations-db [Last accessed 1 February 2023].

Socioeconomic Data and Applications Center (SEDAC). 2023b Data Submission. Available at https://sedac.ciesin.columbia.edu/data-submission [Last accessed 1 February 2023].

Trappler, T. 2009. Is There Such a Thing as Free Software? *The Pros and Cons of Open-Source Software*. Available at https://er.educause.edu/articles/2009/7/is-there-such-a-thing-as-free-software-the-pros-and-cons-of-opensource-software [Last accessed 1 February 2023].

Urquidi Diaz, A. 2021a. Harvestable Metadata Services: A Reference Library. Canada. Available at https://www.zotero.org/groups/2597955/intelibrary_-_harvestable_metadata_services/library [Last accessed 1 February 2023].

Urquidi Diaz, A. 2021b Harvestable Metadata Services Implementation Plan Template. A Guide by the World Data Systems International Technology Office. In: Payne, K, Urquidi Diaz, A and Li, WAW (eds.) *Steps Towards an Harvestable Metadata Services Implementation Plan (flowchart format) & Template: Metadata Service Implementation Plan*. *Zenodo*. Available at https://zenodo.org/record/4589014 [Last accessed 1 February 2023].

Urquidi Diaz, A, Li, WAW and **Payne, K.** 2021. Interactive, step-by-step narrative of Harvestable Metadata Services implementation plan. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.4589036 [Last accessed 1 February 2023].

Urquidi-Diaz, A, Payne, K, Chambodut, A, Chen, RS, Chuang, L, Downs, R, Flower, S, Kun, B, Oyola, MI, Vinay, S, Wang, J, Zhang, Y and **Xu, Q.** 2022. Harvestable Metadata Services Use Case Reports. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.5997733 [Last accessed 1 February 2023].

Valentine, D, Zaslavsky, I, Richard, S, Meier, O, Hudman, G, Peucker-Ehrenbrink, B and **Stocks, K.** 2020. EarthCube Data Discovery Studio: A gateway into geoscience data discovery and exploration with Jupyter notebooks. Concurrency and Computation: Practice and Experience. DOI: https://doi.org/10.1002/cpe.6086 [Last accessed 1 February 2023].

Ventura-Traveset, J, Navarro, V and **Romero, I.** 2019. GSSC Global Data Center Technical Report 2019. In: International GNSS Service: Technical Report 2019. Villiger, A and Dach, R (eds.), *IGS Technical Reports*. Bern, Switzerland: IGS Central Bureau, Bern Open Publishing. pp. 163–166. Available at https://boris.unibe.ch/144003/1/2019_techreport%282%29.pdf [Last accessed 7 February 2023].

Villiger, A and **Dach, R.** 2019a International GNSS Service: Technical Report 2018. *IGS Technical Report*. Bern: IGS Central Bureau, Bern Open Publishing. Available at https://boris.unibe.ch/130408/ [Last accessed 1 February 2023].

Villiger, A and **Dach, R.** (eds.) 2019b. International GNSS Service: Technical Report 2019. *IGS Technical Reports*. Bern, Switzerland: IGS Central Bureau, Bern Open Publishing. Available at https://boris.unibe.ch/144003/1/2019_techreport%282%29.pdf [Last accessed 7 February 2023].

Waide, RB, Brunt, JW and **Servilla, MS.** 2017. Demystifying the landscape of ecological data repositories in the United States. *BioScience*, 67(12): 1044–1051. [Last accessed 1 February 2023]. DOI: https://doi.org/10.1093/biosci/bix117

Wang, J, Bu, K, Wang, Y and **Shao, Y.** 2020. Progress in activities of WDS-China data centers. *Data Science Journal*, 19(1): 33. [Last accessed 1 February 2023]. DOI: http://doi.org/10.5334/dsj-2020-033

WDC-RRE. 2016. The specification of WDC-RRE data identification. Beijing, China: Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences. Available at http://wdcrre.data.ac.cn/page/data_identification [Last accessed 1 February 2023].

WDS Scientific Committee. 2016. World Data System Bylaws. Available at https://worlddatasystem.org/about/bylaws/ [Last accessed 1 February 2023].

Wu, M, Psomopoulos, F, Khalsa, SJ and **de Waard, A.** 2019. Data discovery paradigms: User requirements and recommendations for data repositories. *Data Science Journal*, 18(1): 3. [Last accessed 1 February 2023]. DOI: http://doi.org/10.5334/dsj-2019-003

Yarmey, L and **Baker, KS.** 2013. Towards standardization: A participatory framework for scientific standard-making. *International Journal of Digital Curation*, 8(1): 157–172. [Accessed 1 February 2023]. DOI: https://doi.org/10.2218/ijdc.v8i1.252

Yu, Y, Ibarra, JE, Kumar, K and **Chergarova, V.** 2021. Coevolution of cyberinfrastructure development and scientific progress. *Technovation*, 100: 102180. [Last accessed 1 February 2023]. DOI: https://doi.org/10.1016/j.technovation.2020.102180

Zuo, X, Chen, Y, Ohno-Machado, L and **Xu, H.** 2021. How do we share data in COVID-19 research? A systematic review of COVID-19 datasets in PubMed Central Articles. *Briefings in Bioinformatics*, 22(2): 800–811. [Accessed 1 February 2023]. DOI: https://doi.org/10.1093/bib/bbaa331

TO CITE THIS ARTICLE:
Downs, RR, Urquidi Díaz, A, Xu, Q, Wang, J, Chambodut, A, Liu, C, Flower, S and Payne, K. 2023. Harvestable Metadata Services Development: Analysis of Use Cases from the World Data System. *Data Science Journal*, 22: 20, pp. 1–20. DOI: https://doi.org/10.5334/dsj-2023-020

Submitted: 09 March 2023
Accepted: 13 March 2023
Published: 05 July 2023

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