Exposing Standardization and Consistency Issues in Repository Metadata Requirements for Data Deposition

Jihyun Kim, Elizabeth Yakel, and Ixchel M. Faniel

We examine common and unique metadata requirements and their levels of description, determined by the data deposit forms of 20 repositories in three disciplines—archaeology, quantitative social science, and zoology. The results reveal that requirements relating to Creator, Description, Contributor, Date, Relation, and Location are common, whereas those regarding Publisher and Language are rarely listed across the disciplines. Data-level descriptions are more common than study- and file-level descriptions. The results suggest that repositories should require detailed study-level descriptions and information about data usage licenses and access rights. Moreover, repositories should determine metadata requirements in a standardized and consistent manner.

Introduction

Data repositories use deposit forms and guidelines to declare the transfer of data responsibility as well as establishing the rights and permissions to manage and access the data. The forms and guidelines further serve as tools for collecting metadata and documentation about data to facilitate reuse. Thus, data deposit documentation performs the dual purposes of defining a contract between depositors and repositories and gathering information about the deposited data. Data deposition is a central part of the workflow underlying all data repositories, as well as a major component of data publishing, according to publishing guidelines and models.[1] Few studies have examined data deposition requirements, and even fewer have examined deposit forms' requirements for describing the data.[2] This study compiles and analyzes the information that repositories request during the deposition process in three disciplines: archaeology, quantitative social science, and zoology. Comparing the metadata within and across the chosen disciplines' major data repositories, this study assesses the similarities between them and addresses the levels of description (study, data, and file) the repositories require. Study-level metadata include important contextual information on how the study generated data, such as the study's purpose and data creators. Data-level descriptors concern actual content and

*Jihyun Kim is Associate Professor in the Department of Library and Information Science at Ewha Womans University; email: kim.jh@ewha.ac.kr. Elizabeth Yakel is Senior Associate Dean for Academic Affairs and Professor of Information, School of Information at the University of Michigan; email: yakel@umich.edu. Ixchel M. Faniel is Senior Research Scientist at OCLC Research; email: fanieli@oclc.org.

information associated with data (for example, data collection date and types of data). File-level descriptors identify information about the file, such as the name, format, and version. Documenting data for all three levels is necessary for the sufficient description of the context and characteristics of the data, which helps reusers adequately understand and interpret the data.[3] The specific research questions include the following:

- (RQ1) What metadata elements exist within each discipline's deposition documentation and how consistently are these defined across repositories in a discipline?
- (RQ2) How similar are repositories' data deposition requirements among the three disciplines?
- (RQ3) What levels of description are identified in each data deposition requirement?

Literature Review

Metadata and documentation that describe the content and context of data are essential to promoting data reuse. Collecting and presenting such information in a meaningful manner supports reusers' critical understanding of the data. We begin by reviewing studies discussing reusers' needs for descriptive and contextual information in archaeology, quantitative social science, and zoology. We then examine previous research examining data deposit requirements and metadata and associated content from websites of data repositories. Finally, we end with research on the quality of metadata in data repositories and their influence on data reuse.

1) Data reuse needs for descriptive and contextual information in archaeology, quantitative social science, and zoology

Studies examining the perceptions and experiences of data reusers in archaeology, quantitative social science, and zoology confirm the importance of descriptive and contextual information about data. Archaeologists develop guidelines, standards, and ontologies for data documentation, although they do not fully incorporate the needs of researchers in terms of understanding the context of data creation and collection.[4] Social scientists are satisfied with their data reuse experiences when the included documentation is sufficient, increases their understanding of the data, and is presented clearly.[5] Zoologists emphasize the importance of preserving contextual information about specimens to make sense of the data.[6]

2) Data deposit requirements

Descriptive and contextual information pertaining to data are usually required at the time of deposition. Data deposit forms often include a submission or deposit agreement that specifies the type of data being deposited, metadata, and access restrictions.[7] The data deposit agreement is thus a primary tool for collecting descriptive and contextual information from data producers. Several studies have examined the information required in data deposit agreements and repository websites. Yoon and Tibbo analyze the data deposit requirements of 16 data repositories in the social sciences. They identify and examine in detail three levels of metadata in deposit requirements: the study level, the data level, and the file level.[8] Other studies focus on common and/or unique metadata elements required at submission without specifying the levels of data documentation. Fernel and Shiri examine metadata elements and standards used by four research data services: DataCite, Dataverse Network, Dryad, and Figshare and suggest generalized metadata elements and discuss whether each element is common or unique.[9] Similarly, Assante et al. analyze 11 metadata attributes used by five data repositories: 3TU.Datacentrum, CSIRO Data Portals, Dryad, Figshare, and Zenodo.[10] Austin et al. survey 32 data publishing platforms in various disciplines including archaeology, the life sciences, and the social sciences and identify five common metadata elements: author, publisher, subject, dates of collection, and abstract.[11] In table 1, we compare the metadata elements identified in these studies. The five most common data deposit metadata elements are: creators/contributors, topical subject(s), general description, dates, and methodology.

TABLE 1
Comparison of Data Deposit Requirements in Previous Studies

Fernel and Shiri (2014)	**Yoon and Tibbo (2011)**		**Austin et al. (2015)**	**Assante et al. (2016)**
Titles	Study level	Title of study		
	Data level	Title of data		Minimal description (title)
Creators, contributors	Study level	Principal investigator (co-Principal investigator)	Author	Minimal description (author)
		Agency/funder		Project (funding sources)
		Donor/contact person/ depositor	Readme file (contact information)	
		Data producer (or creator), if different		
			Publisher	
Topical subject(s)	Study level	Subject term	Subject matter	Subject
		Subject/area of investigation		
		Study metadata in general (not specified)		
	Data level	Subject terms for data		
General description	Study level	Description of study	Abstract	Project (research goals, type of research)
	Data level	Description about content included		Minimal description (brief description or abstract)
Object type(s)	Data level	Types of data		
Date(s)	Data level	Data collection date	Dates of collection	Dates of creation, submission and publication
Right, access, use	Study level	Copyright check		License (access rights and licenses)
	Data level	Use of restriction check		

TABLE 1
Comparison of Data Deposit Requirements in Previous Studies

Fernel and Shiri (2014)	Yoon and Tibbo (2011)		Austin et al. (2015)	Assante et al. (2016)
Object technical characteristics	Data level	Data file		
	File level	Data file format		Format (file format including size)
		Document file format		
		Data file size		
		Software name		
		Image file format		
		Audio file format		
		Video file format		
		Data file naming		
		Delivery (media) format		
		Software version		
		File compression		
		Platform		
Spatial subject(s)				Coverage (spatial coverage)
Identifiers	Study level	Identifier		Availability
Temporal subject(s)	Study level	Time period of study		Coverage (temporal coverage)
Citation				Bibliometric data
Versioning	File level	Data file version		
Methodology	Data level	Data collection methodology		Provenance (methodologies, original sources)
		Codebook	Readme file (definitions of column headings & row labels, data codes including missing data and measurement units)	
		Sampling		
		Instrument		Provenance (instrument or software tools)
		Technical information about variables		
		Types and scales of variables		
		Analysis performed on data		
		De-identification		

TABLE 1
Comparison of Data Deposit Requirements in Previous Studies

Fernel and Shiri (2014)	Yoon and Tibbo (2011)		Austin et al. (2015)	Assante et al. (2016)
Methodology	Data level	Data edit/cleaning procedure	Readme file (data processing steps)	
		Data dictionary		
		Relationship between documents/tables/variables		
Related resources	Data level	Final report/publication generated by data		Paper reference
Language(s)				
Status				
Production				
Additional grant information				
Notes				

3) Metadata quality in data repositories

Some studies investigate metadata quality in data repositories and discuss its impact on data reuse. Studies examining the quality of four metadata elements—creator, date, type, and subject—in Dryad note the prevalence of incomplete creator names, inconsistent date format and resource type, and the lack of controlled vocabulary used for subjects. To address these problems, the authors recommend unique creator IDs, standardized formatting, and predetermined lists during metadata entry.[12] Another study evaluates the quality of metadata in HealthData.gov, an open government data repository in the United States. HealthData.gov's metadata requirements include the unique ID, title, description, and URL of each dataset, as well as the author—namely, the federal agency that submitted the data. After submission, repository staff modify selected datasets' metadata to meet their curatorial standards. Marc et al. measured the metadata quality in terms of accuracy, completeness, and consistency, finding the quality of the modified metadata to be significantly higher when compared to unmodified sets of metadata. To improve metadata quality, the study suggests providing mandatory fields, using a standardized vocabulary, following Dublin Core (DC) standards, and updating the metadata entry user interface to offer contextual cues.[13]

Our review of previous studies indicates that, while common descriptive metadata elements are usually required at the time of data deposit, more specific contextual information, such as methodology, is rare. For quantitative social science data reusers, Carlson and Anderson find that numbers and observed raw data are not self-contained; therefore, external information is always needed to explain how data are constructed and manipulated and to help reusers assess the data quality.[14] Yet the studies we reviewed show that these metadata are required infrequently during data deposition. These studies also indicate that metadata in data repositories is often incomplete, inaccurate, and inconsistent in format and terminology.

While Yoon and Tibbo focus on quantitative social science repositories, the other studies focus on descriptive metadata elements that general-purpose repositories require at deposition. Although the study on HealthData.gov concerned the metadata quality of open government data, the results have implications for research data repositories. In addition, there is a dearth of research investigating the gap between information that data sharers are required to submit and information that data reusers need to make sense of the data. We focus on common and unique metadata elements required in data deposit forms within archaeological, quantitative social science, and zoological disciplinary repositories. We also examine how well the deposition requirements support the needs of reusers in each discipline.

Methodology

Our sample encompasses the data deposit requirements in 20 repositories: six in archaeology, eight in zoology, and six in quantitative social science (see table 2). We selected these repositories based on the following criteria: 1) reputation and wide use in their disciplines; 2) published data deposit requirements; 3) significant or solely digital data formats; 4) information presented in English. In some cases, repositories that also manage large physical as well as digital collections were included, such as the American Museum of Natural History (AMNH). Metadata aggregators that harvest and compile data from various sources (such as the Global Biodiversity Information Facility (GBIF) or VertNet) were excluded. The disciplines of quantitative social science, zoology, and archaeology are selected because of their diverse data-sharing and reuse traditions. Quantitative social science has a long data-sharing and reuse tradition, and data reuse is an accepted form of scholarship in many disciplines employing quantitative approaches to social science questions. Repositories holding quantitative social science data are well established with procedural and metadata standards. (For example, the Inter-university Consortium for Political and Social Science has been in existence for more than 50 years.) At the other end of the spectrum is archaeology where data sharing and reuse are in their infancy. Furthermore, there are fewer repositories holding archaeological data, so there are no accepted disciplinary metadata standards in that field; and data reuse is not fully accepted as a form of scholarship and knowledge creation. Zoology falls in the middle. Museums have long held zoological data and specimens, but widespread data sharing online is more recent. While there are disciplinary metadata standards, retrospective conversion of legacy systems to those standards is still in progress and impairs data reuse in museums as well as online.

TABLE 2
Data Repositories Surveyed in the Three Disciplines

Archaeology	Zoology	Quantitative Social Science
• Archeology Data Service (ADS) • The Digital Archaeological Records (tDAR) • Data Archiving and Networked Services (DANS), Archaeology • The Institute for Archaeologists • OpenContext • Parks Canada	• American Museum of Natural History (AMNH) • Canadian Polar Network • DANS, Life Science and Medicine • Dryad • NCBI GenBank • Morphbank • Museum of Vertebrate Zoology(MVZ) • Protein Data Bank	• Australian Data Archive (ADA) • DANS—Social and Behavioral Sciences • ICPSR • Roper Center • The Odum Institute • UK Data Archive

We collected data deposit requirements from the repositories in fall 2017. Most repositories had specific data deposit forms; some, however, such as the Institute for Archaeologists, Dryad, and the Protein Data Bank, did not. In these cases, we extracted data deposit metadata requirements from more generic deposit guidelines. Of note, the Dutch Data Archiving and Networked Services (DANS) uses the same data deposit form for all data deposited regardless of discipline, with specific guidelines for each discipline filling out the form.

After collecting the requirements and identifying the generalized terms, we used two complementary methods of analysis to match the individual metadata requirements from each repository to the generalized terms. First, we determined similarity between requirements based on the definitions, if any, provided in the forms or guidelines. We also assigned a level to the metadata elements: study, data, and/or file. Second, we used the Fuzzy Lookup add-in function of Microsoft Excel to determine similarity of repository terms to DC terms and calculated the index of similarity on a scale from 0 to 1. (An exact match generates a score of 1.) We set the similarity threshold at 0.70, meaning that any match less than 0.70 was deemed inaccurate. The Fuzzy Lookup tool was useful, not only to compare the similarities between terms, but also to understand how repositories categorically describe data. The tool does not, however, consider the similarity of meanings or context and does not detect synonyms.

Findings

This section is organized according to our three research questions. We begin with an analysis examining metadata requirements in each discipline: archaeology, zoology, and quantitative social science (RQ1). Analyzing the metadata required on deposit forms and guidelines, we find that repositories implicitly request generalized metadata that aligned with the 15 DC metadata elements for describing information resources (see table 3). Repositories across the three disciplines also consistently request additional elements, such as Comments and Location. Both archaeological and social science repositories occasionally use Documents and Files fields, while Sequence and Taxon metadata uniquely appear in the forms or guidelines of zoological repositories. We also address the level of data deposition requirements required by each discipline in this section (RQ3). Then we turn to our second research

TABLE 3
Generalized Terms

	Archaeology	Zoology	Quantitative Social Science
DC Metadata Elements	Title	Title	Title
	Creator	Creator	Creator
	Subject	Subject	Subject
	Description	Description	Description
	Publisher	Publisher	Publisher
	Contributor	Contributor	Contributor
	Date	Date	Date
	Type	Type	Type
	Format	Format	Format
	Identifier	Identifier	Identifier
	Source	Source	Source
	Language	Language	Language
	Relation	Relation	Relation
	Coverage	Coverage	Coverage
	Rights	Rights	Rights
Additional Elements	Comments	Comments	Comments
	Location	Location	Location
	Document	Sequence	Confidentiality
	File	Taxon	Document
			File

question and investigate the similarities and differences across disciplines. We end this section with a cross-disciplinary comparison. The findings demonstrate significant inconsistency within and across disciplines, particularly when metadata element definitions are considered.

Data Deposit Requirements in Archaeology

Appendix A presents the similarity scores in parentheses. DANS, using all 15 of the DC elements, has the highest degree of similarity and requires no additional metadata. Similarity decreases from there. tDAR requests nine DC and two non-DC elements; ADS eight DC and one non-DC element; Parks Canada requested six DC and two non-DC elements; and Open Context also asks for six DC and one non-DC element. The Institute for Archaeologists yields only three match results with similarity scores greater than 0.70.

Although the fuzzy lookup procedure identifies exact or similar word matches, it does not consider the meaning behind the terms, resulting in false positives. For example, the DC element Type[15] is defined as "the nature or genre of the resource."[16] This scored a .85 similarity with Parks Canada metadata element "Site: Environment: Soil Type." In Open Context, the "Data format and structures" metadata element receives a .76 match to Date. In addition, Location refers to the physical place names of the places covered by a given set of archaeological data. One match result for this term is "Basic Information: Publisher Location," a requirement of tDAR, which does not reflect the correct meaning. The Institute for Archaeologists' request for "all original written documents created throughout the course of the project" receives a .76 match with Creator. Considering the Institute's intent of this result, this should have matched the generalized term Document, yet the result that matches Document is "the schedule of works or similar documents."

Due to the limitations of Fuzzy Lookup, we perform a qualitative content analysis to determine which data deposit requirements are meaningfully related to DC. Based on the analysis, the number of archaeological repositories requesting data depositors to enter each metadata element is presented in table 4. The most common requirements are Description and Rights, required by all six repositories. The second most common elements are Title, Creator, Subject, Contributor, Relation, Coverage, and Location. Four repositories provide Date, Type, Format, Identifier, and Comments. Three repositories ask for Document. Only two repositories require Publisher, Source, Language, and File, a set of elements uniquely required by archaeological repositories.

In addition, we examine the requirements requesting different levels of description (see table 4). The level of description that each repository defined for data deposit requirements usually meets one of three levels: study level, data level, or file level. In some cases, however, one requirement encompasses both study and data levels of description. For example, tDAR requires "General keyword(s)" at both the study- and data-level description. In this regard, table 4 provides columns for study level only, data level only, and both study and data levels to distinguish the repositories that identify a requirement defined exclusively as study level or data level from those that list a requirement that covers both study and data levels. The same distinctions are made in tables 5 and 6.

Repository deposit guidelines and definitions on the data deposit forms provide a more nuanced understanding of the metadata required. Even when the intent of a term is consistent across repositories, terminological differences still occur. All subject repositories require information paralleling Description and Rights elements. Further, all ask depositors for a study-level

TABLE 4
The Number of Repositories in Archeology that Determine Data Deposit Requirements at Different Levels of Description

Generalized Terms		Repositories that Determine the Requirement	Levels of Description				Notes
			Study Level Only	Data Level Only	Both Study and Data Levels	File Level	
DC Core Elements	Title	5	2	4			tDAR requires "project name" (study level only) and "title" (data level only)
	Creator	5	5				
	Subject	5	2	3	1		tDAR requires "Investigation type(s): keywords" (study level only) and "General keywords" (both study and data levels)
	Description	6	5	3	1		tDAR requires "Basic information: Description," which covers both study and data levels
	Publisher	2		2			
	Contributor	5	5				
	Date	4		2	2		"Project dates" (ADS) and "Date" (DANS) cover both study and data levels
	Type	4		4			
	Format	4				4	
	Identifier	4	1	2	2		"Item-Specific or Agency Identifiers" (tDAR) and "Identifier" (DANS) cover both study and data levels
	Source	2		2			
	Language	2		2			
	Relation	5		5			
	Coverage	5	1	5			Parks Canada identifies several requirements regarding coverage at both study level only and data level only.
	Rights	6	1	5			
Additional Elements (Non-DC)	Comments	4	1	3			
	File	2				2	
	Document	3		3			
	Location	5		5			

description. The element names used, however, varied: tDAR and DANS employ "description." Open Context uses "abstract" and ADS "introduction"; both also ask for the image of a research project's website or banner in the study description. Parks Canada lists "comments" for a site, which is defined as a study-level description and also requires several other study-level description elements (for example, "rationale for lot" or "excavation method"). The Institute for Archaeologists requests "project specification or research design" as a study-level description. Interestingly, Open Context is the only repository that requires methodology-related information, including "clean and edits" and "decoding" at the data level, although elements describing the data are required by three repositories. ADS asks for an "overview," and Open Context a "short description." tDAR and Open Context specify elements for describing the current disposition or status of data, defined as "a piece of administrative metadata that controls the resource's status within the archive" (tDAR). Three repositories request information concerning Document. The Institute for Archaeologists provides a detailed list of required documentation, including "the schedule of the work," "all original written documents," or "all original drawings." Open Context seeks "methodological notes" and Parks Canada asks depositors to submit references to the documentation (for example, a "field notebook reference"). By requesting various pieces of evidentiary information, these metadata requirements use varied approaches to better situate studies and associated data in a more cohesive documentary framework.

Rights is the other DC requirement mentioned by all archaeological repositories. Parks Canada is the only archaeological repository that determines rights-related elements at the study level, including elements about site ownership and the legal description of the site. The remaining five repositories ask for information about copyrights or copyright holders (ADS, DANS, and the Institute for Archaeologists), licenses and license holders (ADS, DANS, and Open Context), and access rights to the data (tDAR and DANS).

Five repositories require Title, Creator, Subject, Contributor, Relation, Coverage, and Location. tDAR and Parks Canada ask depositors to indicate a title at a study level (that is, "project name"). Four repositories require Title at the data level, defined as a title or name for the dataset (for example, "a short descriptive name of the dataset" for Open Context). tDAR is the only repository requiring Title at both study and data levels. Forms also frequently seek Subject-related metadata, most often using the terms "subject" or "keywords." As a study-level description, tDAR asks depositors to submit "investigation type(s): keywords," defined in its data dictionary as 23 types of archaeological investigation (for example, "architectural survey" or "site evaluation/testing"). DANS also requires Subject at the study level requesting that it be "used to describe the contents of the research project." tDAR asks for "General keyword(s)" at both study- and data-level descriptions. Three repositories—ADS, Open Context and Parks Canada—ask for Subject only at the data level. ADS specify this as "keywords for the subject content of the dataset." Parks Canada requests Subject at the data level in the form of images or other media.

Several metadata elements indicate roles in data production. DC element Creator primarily manifests as "data creator," although Parks Canada uses different terms, including "researcher" and "staff name(s)." tDAR disambiguates "author/creator" into "person" and "institution" and provides a list of various "individual and institutional roles," some of which are relevant to Creator, such as principal investigator or project director. Forms use Contributor to indicate individuals or institutions in various roles. ADS and Open Context ask depositors for grant "project funders" and institutional "support."

tDAR identifies a number of "individual and institutional roles." One is "contributor," "an individual who contributes to but is not the primary creator of" data. The roles also include sponsors and landowners. DANS defines "contributor" as persons or organizations that have written a part of the publication, collected, or recorded data. Parks Canada cites several types of contributors, including "informant name"—a person who has special knowledge about a site—and data "recorders."

DC defines Relation as a related resource. ADS follows this definition for its "related resources" field (ADS). Open Context asks more narrowly for "related publications," while Parks Canada asks for the relation between data and an archaeological site, events that happened, or objects that are discovered at a given site, including bibliographic references about each site.

DC Metadata Initiative defines Coverage as the "spatial or temporal topic of a resource."[17] We draw distinction between spatial and temporal coverage because the data deposit guidelines and forms do so in practice. In our analyses, we use Location to indicate spatial aspects and Coverage temporal metadata. In archaeology, descriptions of Coverage involved cultural periods or radiocarbon periods of data. Location is called upon to describe the positioning of archaeological data collection sites. ADS and Open Context require one or two elements termed "location" or "site name." tDAR and DANS require elements relating to the site (for example, "site type" [tDAR] or "spatial point" [DANS]). Parks Canada seeks more detailed elements regarding the site, such as fields to identify geographical coordinates, cultural region, province, district, or township.

Four repositories mention Date, Type, Format, Identifier, and Comments. Although the request of Date usually references the data creation date, its meaning differs widely across the repositories. ADS asks for "project date," including the date of data creation, the start and end dates of a research project, the data processing date, and the computerization date at the study or data level. Similarly, DANS requires a "date," the "date on which the research project is finished and the dataset is completed." At the data level, tDAR specifies "year created," and "file information: date," which suggest the date when a submitted file is created. Parks Canada also identifies data-level description for Date. The requirement at the study level concerns the range of dates when the data gatherer visits a site or units of the site. A data-level Date field indicates when the data are generated. Further, DANS asks for "date available," which allows depositors to impose temporal release restrictions on their data.

Type indicates either the data's media format or the material types of objects collected from an archaeological site. ADS' label is "data type," indicating media, such as documents, images, or video/audio. tDAR provides a list of "material types" that can be found in an archaeological site (for example, "ceramic," "chipped stone," or "fauna"). DANS defines "type" as a "general characteristic of dataset" and suggests that it can be used for both the archaeological and technical descriptions of data contents. Similarly, Parks Canada suggests elements related to either material type or media type. The "objects: samples" element indicates material types discovered at a site—for example, stone tools or bone. The "image/recordings" element provides various options of media, including "digital image (still)" or "air photo." Type as media is considered format by other repositories. ADS provides "preferred and accepted file formats," which correspond to each data type accepted by the repository. For example, spreadsheets are a data type, listing preferred file formats including .csv, .xls/.xlsx, and .ods. DANS asks depositors to submit not only a format, but also software information, such as the manufacturer, the product name, and the version number. For Parks Canada, "format" indicates only whether the data are digital or analog. Open Context prefers data for ingest in tabular format, such as .csv., and the abstract or background of a project in MS-Word or a similar format.

Unique Resource Identifiers are common for published resources, but less so for data. tDAR and DANS ask for identifier information at both the study and data levels. "Item specific or agency identifiers" determined by tDAR indicate the "name of any agency or project identifier" and "a list of the specific identifiers known to the resources." DANS wants "identifiers" "to make it possible to identify the research project or the files." ADS requires "identifiers" "specific to the collection," which indicate a data-level description. Parks Canada requires various kinds of identifiers at only the study level, such as unique identifiers and/or units of a site. At the data level, various identifiers for image or multimedia data are specified, such as "pages," "roll numbers," or "reel numbers." Some repositories account for miscellaneous information in a Comments field variously specified as "notes" (tDAR), "remarks" (DANS), or "potential application of data" (Open Context).

Unique archeological data deposit requirements include Publisher, Source, Language, and File. Publisher is required by tDAR, listing it as "document publisher," whereas DANS lists it as "the organization that published the dataset or the publication." tDAR also uses Source, meaning "source collection" drawn from a published or unpublished work. DANS also requests information about the "source" "on which the digital data in the dataset have been based." Language is mentioned by ADS and DANS, and both repositories define it as the language in which data and related documents are written. Finally, File is defined in a greater detail by ADS than other repositories. ADS provides metadata templates for files in different formats (for example, spreadsheets, GIS, or images).

Data Deposit Requirements in Zoology

Fuzzy Lookup results suggest that the zoological repositories, except in the case of DANS, sparsely use generalized terms for the data deposit requirements (see appendix B). Because DANS requires the same metadata at the stage of data deposit across disciplines, all 15 DC elements match the data deposit requirements exactly or approximately. Morphbank asks for metadata that match eight generalized terms, including seven DC and one non-DC element. NCBI GenBank has four elements that match five generalized terms. Similarly, Dryad has three data deposit requirements that match five generalized terms; one metadata element, "Description for each file or group of files: type(s) of data included," match three generalized terms: Description, Type, and File. The remaining four repositories have match results from 1 to 3 generalized terms. While seven repositories use Date, there are two false results—"Data access (CPN)" and "Contact information for author(s) regarding data analyses (Dryad)." Further, Description and Type are used by four repositories respectively, and Source is used by three. Otherwise, terms are used by one or two repositories. Comments is not used at all.

Irrelevant matches are generated in the Fuzzy Lookup process for zoological repositories. "Date of curator's signature (AMNH)" is returned as an approximate match for Creator. In addition, "Plain language summary (CPN)" is matched to Language, rather than Description. Publisher matched to "Citation(s) of your published research derives from these data" (Dryad) and "Image: Date to publish" (Morphbank). Format proves problematic, such as the match with "Specimen: form" (Morphbank). Source is deemed similar to "Organism name, applicable source modifiers, location" (NCBI GenBank), although source modifiers describe information about organisms from which researchers obtain nucleic acid sequences.

Our content analysis of metadata definitions in the zoological repositories provides more meaningful comparisons among terms. Table 5 presents the frequencies of requested metadata elements and aligns these with DC terms at the different levels of description. Out of eight repositories, seven ask for Description and six seek elements related to Creator and Contributor. Five repositories request Date, five request Location, and four request Relation. Thus, the data deposit requirements relating to these six generalized terms—Description, Creator, Contributor, Date, Location, and Relation—are mentioned by a majority of the repositories. Among the remaining 14 generalized terms, Title is required by three and Subject by two repositories. Several terms relating to data deposit requirements are only listed by one repository, including Publisher, Format, Source, Language, File, and Sequence. All but File and Sequence are required by DANS, the only repository that uses all DC elements. The data deposit requirements relevant to Publisher, Format, Source, and Language are rarely listed by zoological repositories.

TABLE 5
The Number of Repositories in Zoology that Determine the Data Deposit Requirements at Different Levels of Description

Generalized Terms		Repositories that Determine the Requirement	Levels of Description				Notes
			Study Level Only	Data Level Only	Both Study and Data Levels	File Level	
DC Core Elements	Title	3		3			
	Creator	6	6				
	Subject	2	1	1			
	Description	7	2	6			
	Publisher	1		1			
	Contributor	6	6				
	Date	5		5	1		DANS requires "Date" covering both study and data levels
	Type	2		2			
	Format	1				1	
	Identifier	2		1	1		DANS requires "Identifier" covering both study and data levels
	Source	1		1			
	Language	1		1			
	Relation	4		4			
	Coverage	2		2			
	Rights	3		3			
Additional Elements (Non-DC)	Comments	2		2			
	File	1				1	
	Location	5		5			
	Sequence	1		1			
	Taxon	3		3			

With respect to levels of description, metadata elements relating to Subject and Description can pertain either to the study or data level. DANS is the only repository that asked for Date and Identifier at both the study and data levels. DANS and Dryad seek file-level descriptions for Format and File, respectively. All the zoological repositories require Creator and Contributor, which involve the study level. Other than these cases, data deposit metadata in the zoological repositories is confined to the data level.

Description is the most frequently requested metadata element. Two repositories—CPN and DANS—require study-level descriptions. CPN ask for "purpose," "abstract," and "summary" of research projects. CPN's description also encompasses the methodologies for collecting or processing data and the "status of data collection/production." DANS requires Description concerning the background of a study, such as "a description of the character and the purpose of the research project, the nature of the data, and the most significant conclusions."

In all, six repositories have data-level requirements for Description. In particular, AMNH, MVZ, and Morphbank require the number of specimens or items. Various characteristics of specimens are also required in Description, such as the form and developmental stage (Morphbank), or the class, condition, and data quality (MVZ). Protein Data Bank requires "three-dimensional atomic coordinates" of proteins and "information about the composition of the structure." Two repositories, MVZ and CPN, ask for either the physical or digital location where the data exist. MVZ requests an element named "MVZ location," indicating physical storage within the museum; CPN asks for "links to data" or, if these are not available, required the email address of a principal investigator. In addition, AMNH requests an element relating to administration, namely whether a specimen is a gift, exchange, or purchase. Methodological information is also required at the data level. Data processing or specimen preparation information is required by four repositories. For example, Morphbank asks depositors to describe "specimen preparation type" and "imaging technique." Dryad also requests information about de-identification procedures "for sensitive human subjects or endangered species data," as well as the platform and software used for analyses.

In zoology, Taxon, requested by three repositories, is related to Description. Morphbank wants the most detail: "type of name," "rank identification," "name source," "name status," "vernacular," and other details. GenBank and MVZ ask for Taxon by simply requiring the name of an organism or specimen. GenBank also asks depositors to provide the specific species name (if known) and the best taxonomical name (if new or unrecognized).

Creator or Contributor are required by six of the zoological repositories. With respect to Creator, all repositories—except for the two museums (AMNH and MVZ)—have this requirement. DANS is the only repository to use the term "Creator" by name. Dryad, NCBI GenBank, and Protein Data Bank use "author(s)" and Dryad and Protein Data Bank require contact information. Morphbank uses the term "Contributor" by name, defined as the "person having the authorization to release the images" of specimens. This concept of a Contributor is more likely to describe a Creator, since creators are responsible for producing and releasing data. Additionally, CPN provides a "responsible parties" section in which to record the names of a principal investigator and one originator. Contributor signifies an even more diverse set of roles. AMNH and MVZ require "donor information," including the name, address, phone, and email; CPN for names of "collaborator(s)" as part of its "responsible parties" section; and Morphbank for the "name(s) of person(s) who determined the taxonomic category of the specimen" and "the name(s) of the collector(s) responsible for collection of the specimen or taking the observation." GenBank also requires the name and contact information of the "submitter."

Overall, five repositories request Date and Location. Different kinds of dates are required, however: the date when specimens are collected or found (AMNH, MVZ, Morphbank); when the study is released or published (NCBI GenBank, Morphbank); when specimens are received by the repository (AMNH, MVZ); when specimens are identified (Morphbank); and the dates of donors' or curators' signatures on the deposit form (AMNH). As mentioned in the section of data deposit requirements in archaeology, DANS has the same Date requirements, including the date available after an embargo period and the dates of the study- and data-level descriptions.

AMNH and MVZ ask for a high-level location where specimens are found, such as the county or state, and CPN requires the name of the "study site" as it is listed in the Canadian Geographical Names Database, which provides official geographic names for locations in Canada. CPN also asks depositors to specify the "research area" based on geographic coordinates. Similarly, DANS requires "spatial coverage" for the name or coordinates of a geographic location, while Morphbank requests a range of elements relating to the "locality" of the specimen, including country, latitude and longitude, minimum and maximum elevations and depths, and a narrative description of the locality. GenBank additionally asks for "source modifiers," which describe information about organisms, such as where a sequence is obtained and "location" presented as an organelle, a specialized structure within a cell, or a map and/or chromosome.

Four repositories request Relation. Dryad requires citations of publications mentioning deposited data; descriptions of associated datasets; and the relationship of the submitted data to the tables, figures, and sections within the publication. Regarding specimens, Morphbank asks for external links and references to the locality, the "relationship type," and "related catalog items" of each specimen, as well as an image depicting each. Morphbank includes a separate metadata form for related publications, while DANS requests the Relation to publications, reports, websites, or other resources related to data.

Three repositories ask for Title and Rights. In terms of Title, CPN, DANS, and Dryad requires data-level description—"title of the data" (CPN), "title" (DANS), or "project name" (Dryad). Rights information is requested by AMNH to confirm whether the collection of specimens inside or outside is conducted legally, while CPN asks for "rights" about "data access" and provides six options: (1) public, (2) limited, (3) limited due to human subject issues, (4) limited due to intellectual property issues relating to local or traditional knowledge, (5) limited due to harm to the environment or the public, and (6) limited due to the use of pre-existing data subject to access restrictions. Finally, DANS requires information about rights holders, license holders, and access rights.

Two of the zoological repositories ask for the following elements: Subject, Type, Identifier, Coverage, and Comments. Each element is requested by DANS and another repository (CPN for Subject and Coverage; Dryad for Type; Morphbank for Identifier and Comments). CPN seeks a Subject and at least five "keywords" associated with the data. Dryad requests the "type(s) of data included" in each file and Morphbank requires several identifiers related to specimen, such as "institution code," "collection code," and "catalog number." In terms of Coverage, CPN identifies "time period"—meaning the time of data collection or the time period the data covered. Comments are requested by Morphbank specifically as "determination notes" and "notes" for each specimen.

As mentioned earlier, only DANS requires Publisher, Format, Source, and Language. In addition, Dryad is the only repository that requires File information and asks for the list and descriptions of the files deposited. Only one repository mentioned sequencing information—GenBank—which specifies molecule type, typology, and length of nucleotide sequence(s). GenBank also requires the file submission or the use of an input form to describe "features of the sequence."

Data Deposit Requirements in Quantitative Social Science

We conducted a Fuzzy Lookup comparison of data deposit metadata with six quantitative social science data repositories. All except DANS used 10 or fewer of the 20 generalized terms (see appendix C). The Odum Institute produces 11 matches when the "description of data content" matches with both Description and Comments, although the match with Comments is incorrect since the fuzzy lookup aligned "content" and Comments. Discounting this result, the Odum Institute uses 10 generalized terms, with eight DC and two non-DC elements. The UK Data Archive (UKDA) has 9 matches with the generalized terms—five DC and four non-DC, with one false positive linking Date to "data collectors." ICPSR uses eight generalized terms—six DC and two non-DC elements. The Australian Data Archive (ADA) and the Roper Center each had eight match results. However, ADA has one false positive. For the Roper Center, both Type and Coverage are matched with "sample type (including geographic coverage)." In addition, the following match for Location is returned: "Location of the weights in the study and a description of the weighting factors." This does not correctly reflect the meaning of the generalized term Location, which we define as a physical place where data collection or other relevant activities are conducted. Therefore, ADA uses seven generalized terms, with five DC and two non-DC elements, and the Roper Center uses six generalized terms, including five DC and one non-DC.

Several generalized terms are frequently used by the repositories (see appendix C). Title, Date, and Coverage are used by all six. Five repositories use Description and File. Type, Format, and Source are used by four repositories. Three repositories use Subject and Location. The remaining terms are used by one or two repositories, or not at all. Two repositories use Creator, Confidentiality, and Document. The terms Publisher, Contributor, Identifier, Language, Relation, and Rights are used only by DANS. No repositories use Comments.

Table 6 presents the number of social science repositories that identify data deposit requirements, as identified through qualitative content analysis. Repositories in the quantitative social sciences demonstrate the highest degree of similarity in data deposit metadata of all the disciplines we study. Almost all repositories require Creator and Contributor, which concern the study level. Data deposit metadata relating to Title, Subject, and Description can be at the study or data levels. However, DANS is the only repository to specifically request Date and Identifier at both study and data levels. Most other repositories focus primarily on collecting metadata at the data level only (see table 6). Only two elements are required at the file-level format, which four repositories seek, and file information, which is wanted by five.

All six repositories in the social sciences require Title, Description, Contributor, Date, and Relation. Three repositories, including ADA, the Roper Center, and UKDA request study-level titles, such as "Study title" (ADA) or "Title of the survey" (the Roper Center). UKDA uses the term "title" explaining that it "should reflect the nature and subject of the data collection and include a date, e.g., *General Household Survey, 2001–2002*." The remaining three repositories focus more on the data; ICPSR asks for "Title of the data collection" (ICPSR) and Odum a "Descriptive title of the data."

TABLE 6

The Number of Repositories in Quantitative Social Science that Delineate Data Deposit Requirements at Different Levels of Description

Generalized Terms		Repositories that Determine the Requirement	Levels of Description				Notes
			Study Level Only	Data Level Only	Both Study and Data Levels	File Level	
DC Core Elements	Title	6	3	3			
	Creator	5	5				
	Subject	4	2	3			UKDA required "Subject categories" (study level only) and "Main topics" (data level only)
	Description	6	4	5			
	Publisher	1		1			
	Contributor	6	6				
	Date	6		5	1		DANS requires "Date" covering both study and data levels
	Type	4		4			
	Format	4				4	
	Identifier	1			1		DANS requires "Identifier" covering both study and data levels
	Source	3		3			
	Language	1		1			
	Relation	6		6			
	Coverage	5		5			
	Rights	4		4			
Additional Elements (Non-DC)	Comments	2		2			
	Confidentiality	3		3			
	Document	4		4			
	File	5				5	
	Location	5		5			

Four repositories (ADA, DANS, ICPSR, and UKDA) request Description at the study level and specify a description or the abstract of a study. Five repositories seek a description of the data and/or information about methodologies. For example, the Odum Institute asks for a "description of data contents." In addition, all repositories except for DANS require precise methodological notes. The most common requirements for the methodology include sampling procedures and method of data collection, which all five repositories request. Population and information about weighting are also common requirements, which four repositories seek.

Three repositories ask for the response rate and two for the unit of analysis. ADA, ICPSR, and the Roper Center list 10–16 options for the method of data collection. ADA also provides 13 options for sampling procedures. In addition, unique elements related to methodologies existed: (1) “Actions to minimize losses” (ADA); (2) the URL of a website that contained information relevant to the data collection (UKDA); (3) “Measurement tools/Scales” and “whether the data collection is one of a series or not” (ICPSR); (4) “Sample type (include geographic coverage),” “Sample description,” and “Size of sample,” (Roper); and (5) “Study design and methodology statement,” “Measurements of sampling error,” “Eligibility criteria,” “How many distinctly different samples are included,” whether “the multiple datasets can be separately used for analysis,” and “procedures for data cleaning” (Odum). Social science repositories also seek documentation to enhance Description. DANS provides the list of documentation specifically for data in the social and behavioral sciences, such as the following: (1) questionnaires or other research instruments; (2) a fieldwork report, if available; (3) a codebook or description of variables and information about methodologies; and (4) publications based on the data or their citations. ADA, the Roper Center, and the Odum Institute also require questionnaires and codebooks. The Odum Institute defines a codebook as any materials that helped in secondary data analysis. ADA additionally requires user guides and technical reports.

Finally, all repositories require publications related to the deposited data, so Relation is an important type of information complementing description. ADA and ICPSR request “related publications” only, whereas ADA specifically asks for citations, links, and descriptions of the publications. The Roper Center and the Odum Institute seek reports in addition to publications, including articles or press releases. Aside from reports and publications, UKDA asks about “related data,” and DANS requests websites and other resources related to the data.

All repositories ask for some type of Contributor. Five repositories are interested in a sponsor or funding agency. Four repositories ask for the name of the depositor; three request the data collector’s name. ADA and UKDA provide fields for “other acknowledgements.” These other roles associated with the Contributor include the “research initiator” (ADA), “grant manager” (ICPSR), “data producer” (the Odum Institute), and “contact person” (the Odum Institute). The Odum Institute also provides an example of a “data producer” as an “organization responsible for bringing the data to its final computerized form.” Both “data producer” and “contact person” are required only if they are different from the data depositor. Furthermore, ADA’s “research initiator” requirement means a particular person or organization for which a study is conducted. Related to Contributor, five repositories ask for a Creator. Three repositories specify that this is the “principal investigator.” For example, DANS and UKDA ask for “creator” and “data creator,” respectively, and explain that principal investigators should be mentioned first for the requirement.

All repositories request Date, generally identified as the data collection date. UKDA seeks “dates of fieldwork.” DANS asks for the date on which the research project and the datasets are completed, as well as a “date available” after an embargo period. Five of the social science repositories request Coverage, File, and Location. Coverage is used to refer to the time period covered by a study, and all repositories except DANS use the term “time period”—for example, “study time period” (ICPSR). DANS requires “temporal coverage,” which indicates “the period of time to which the data relate (date) recorded.” Thus, the Coverage and Date metadata elements’ definitions overlap in some cases. Two repositories, ADA and UKDA, request the “time dimension,” or whether the data collection took place at one point in time or at more than one point in time. The two repositories provided prompts for the time dimension, such as a one-time cross-sectional study, a longitudinal study, or a time series.

File-level information is sought in a variety of different ways. ADA provides a separate section in the deposit form for "Data file description." The section includes two parts for "quantitative data files" and "qualitative data files." For "quantitative data files," the following requirements are included: (1) "file information," such as "file name" and "file content"; (2) "sample description," such as the size of original sample; and (3) "weighting," such as whether the file is weighted as well as the names of the weight variables. The requirements for "qualitative data files" are "file information" and "brief description of the file." While ADA requests precisely detailed descriptions of data files, the other four repositories require minimal information. The Odum Institute and UKDA additionally ask for the edition version of each file, and the Odum Institute requests the "file layout" if the raw format of a file is deposited.

Location commonly refers to "geographic coverage" (4 repositories), or "spatial coverage" in the case of DANS. ICPSR and UKDA request "geographic unit" and "spatial units," respectively. UKDA identifies various dimensions for these spatial units, including "administrative," "postcodes," or "census geography." ICPSR lists five geographic units, "census tract," "state," "precinct," "country," and "other" and specifies selection of the smallest unit that can be analyzed. Source is specified by three repositories and can be akin to Location, Relation, or Creator. DANS, The Odum Institute, and UKDA require "source(s)" of data, which can include other data or printed resources (Relation). UKDA also asks for information about "source location and access (contributor/other role)."

Four repositories in the social sciences request Subject, Type, Format, Rights, and Document. DANS and UKDA require study-level Subject descriptors. In fact, UKDA lists 23 Subject categories and asks that up to six categories be selected. ICPSR, the Odum Institute, and UKDA also ask for "keywords" (ICPSR) or "main topic" (UKDA) at the data level. DANS and the Odum Institute ask the depositor to describe the Type of data in free text, whereas ADA (32 options) and ICSPR (16 options) provide prompts to guide selection. Both repositories list "quantitative," "qualitative," "experimental data," "observational data," "survey data," "census data," "administrative records," and "clinical data." The Roper Center and the Odum Institute inquire about format including ASCII and statistical software file formats, including SPSS, SAS, and Stata. The Odum Institute asks that data layout descriptions in ASCII be included if the data file has a raw format; it also provides MS Access and MS Excel file formats as additional options. ADA requests file formats for quantitative and/or qualitative data files as well as for "deposited documentation, publications and reports."

Rights metadata encompasses both copyright and access rights. DANS requests information about both "rights holders" and "access rights," whereas ICPSR asks whether copyrighted materials requiring special consideration are included in the deposited data. The Roper Center and the Odum Institute inquire about any extant restrictions or embargos for newly deposited data. Repositories also seek restriction information through a Confidentiality metadata label. ICPSR asks whether "confidential information" exists in the deposited data. The Odum Institute inquires whether data "de-identification" has occurred and, if so, requires "the process of de-identification and rules/decisions made." Similarly, UKDA wants to know if the data are "anonymized" and requests the description of "any confidentiality/anonymization issues" relating to the data.

Two repositories in the social sciences specified allow Comments. ICPSR seeks "additional information," and the Odum Institute asks "what else one should know about these data or the study." Only DANS requires Publisher, Identifier, and Language.

Comparing the Data Deposit Requirements across Disciplines

After examining the metadata requested at deposit within each discipline, we analyze commonalities and unique elements across disciplines based on the percentage of repositories specifying each type of metadata element (see table 7). Several elements are similar in archaeology and the social sciences, including Title, Subject, Coverage, Type, Format, Rights, and Document. As previously mentioned and demonstrated in table 5, zoological repositories exhibits fewer common requirements than repositories in the other two disciplines. In addition to the common requirements, several unique elements are identified. Publisher and Language are more prevalent in archaeology. Furthermore, various types of Identifiers are requested by more than half of the archaeological repositories, but rarely by repositories in the other disciplines. Similarly, the Comments requirement is used heavily in archaeology but rarely in both zoology and social sciences. In social science repositories, Source and File are common but not in the other disciplines. The Confidentiality requirement also appears only in social science repositories. Taxon and Sequence are likewise unique in zoological repositories, though not universal. In addition to individual metadata elements, social science repositories provide options to add a considerable amount of detail about methodologies. This is not the case for the other disciplines.

TABLE 7
Common and Unique Data Deposit Requirements within Each Discipline and across Disciplines

	Archaeology (6 repositories)	Zoology (8 repositories)	Quantitative Social Science (6 repositories)
All repositories	Description, Rights		Title, Description, Contributor, Date, Relation
75%–99% of the repositories	Title, Creator, Subject, Contributor, Relation, Coverage	Creator, Description, Contributor	Creator, Coverage, File, Location
50%–74%	Date, Type, Format, Identifier, Comments, Document, Location	Date, Relation, Location	Subject, Type, Format, Source, Rights, Confidentiality, Document
25%–49%	Publisher, Source, Language, File	Title, Subject, Identifier, Coverage, Rights, Comments, Taxon	Comments
24% and less		Publisher, Type, Format, Source, Language, File, Sequence	Publisher, Identifier, Language

Discussion

In the previous section, we addressed our three research questions. We identified which metadata elements exist within each discipline's deposition documentation and showed the degree to which repository definitions of some elements differ. We also pointed to the similarities and differences between the data deposition requirements in the three disciplines as well as differences in what metadata elements are applied at each of the three levels of description. Overall, we find substantial differences in data deposition requirements and the applications of these requirements within and between disciplines. In this discussion, we discuss three salient implications for data curation based on the answers to our research questions and relate these to the FAIR guiding principles: (1) need for sufficient information about context; (2) use of standardized vocabulary to support interoperability; and (3) need for specific information about rights. The FAIR guiding principles also identify these implications as important characteristics of metadata to make data findable, accessible, interoperable, and reusable. Under these principle, the three tenets are these: (1) (meta)data should be released with a clear and accessible data usage license; (2) (meta)data require provenance; and (3) (meta)data need to meet domain-relevant community standards.[18] In particular, FAIR advocates that, to make data reusable, they must be richly described with accurate and relevant attributes.

Need for sufficient information about context

Studies and guidelines focusing on the FAIR principles suggest two types of metadata required for making data reusable: (1) intrinsic and (2) user-centric (submitter-defined or user-expanded) metadata. Intrinsic metadata are "factual information that is 'indisputable' about the data object,"[19] such as the time of data collection, the creator, rights, and the instrumentation used to generate the raw data. Some intrinsic metadata can be captured automatically through metadata extraction. User-defined or user-expanded metadata are "any needed information to properly (re)use the data"[20] and provide a rich description of "the context under which data was generated."[21] User-defined metadata can be added by a data creator or possibly by reusers, who might contribute information on errors or bias they identified during reuse.[22] Specific types of intrinsic and/or user-defined metadata lack clear definitions.[23] Our findings contribute to an understanding of what types of intrinsic and user-defined metadata repository staff collect and how we can provide guidance for improving the metadata. The findings also point to the importance of understanding how metadata at different levels (study, data, file) contribute to an understanding of both content and context and what metadata researchers should collect at what level. The necessity of specifying metadata definitions is key, as the same metadata elements appear at different levels. For example, a study-level metadata description would provide background (the purpose of collecting data, lab conditions) while a data-level description would provide data characteristics (missing data, weighting, codebooks, and the like). Our findings show that Description about methodology is considered at both study and data levels by various repositories, which is problematic when trying to aggregate metadata by linking between levels.

Our findings demonstrate that data deposit metadata varied between archaeology, zoology, and quantitative social science. In particular, the breadth and depth of contextual information required by the repositories vary across disciplines. For example, all but one of the social science repositories require detailed methodological information but only half of the zoological and one of the archaeological repositories require this type of information.

Disciplinary differences also are visible in the various metadata labels and roles specified around creators of and contributors to data. Overall, repositories in archaeology and quantitative social science tend to cover several requirements relating to context, whereas the zoological repositories do not.

Disciplinary communities are developing unique metadata regimes to meet domain-relevant community standards for contextual and content-related information. Considerable variation, however, remains within disciplinary repositories, particularly in archaeology and zoology. In the former, this may be attributed to the newness of the repositories and the lack of a disciplinary culture of data sharing and reuse.[24] In zoology, the variation is due to the broad range of repositories from traditional museums to repositories holding biological entities (GenBank and Morphbank). Still, when like metadata elements are requested, the definitions are often different or the labels for the different elements with like definitions can be different. This leads us to our second implication concerning standard vocabularies:

Use of standards-based vocabulary to support interoperability.

Interoperability is one of the FAIR principles that can be realized if metadata use "formal, accessible, shared and broadly applicable" language and/or vocabulary as well as include "qualified references to other (meta)data."[25] FAIR advocates for domain-relevant community standards and encourages the use of a common template and vocabulary for metadata and documentation. Willis, Greenberg, and White also emphasize the interoperability of a discipline-specific metadata scheme for data to support interdisciplinary research in contemporary science.[26]

Some of the repositories examined in this study recommend the use of thesauri or controlled vocabularies in their data deposit requirements, particularly for subject or temporal coverage. DANS provides a pull-down menu of standardized subject terms for archaeologists, namely the Archaeological Basic Register (ABR). The ABR enables depositors to choose descriptive keywords and site types, though the terms are all in Dutch. DANS also offers a free-text option to describe subject terms or keywords for depositors in other disciplines. In terms of Coverage, DANS uses standardized terms for the time periods determined by ABS for archaeologists, in addition to providing a box for entering free text to describe temporal coverage. Three other archaeological repositories (ADS, Open Context, and Parks Canada) also recommend controlled vocabularies in their data deposit forms or guidelines. ADS suggests that depositors use the UK archaeological thesauri listed on the Heritage Data website for describing Subject. Open Context explains that they use controlled vocabulary with Uniform Resources Identifiers (URIs) from the Library of Congress or link researchers' definitions of biological taxonomy to the most equivalent URIs from the Encyclopedia of Life. Parks Canada recommends the use of (unspecified) controlled vocabularies when describing archaeological objects and the nonmovable features of a site. The findings indicate that repositories for archaeology generally encouraged the use of controlled vocabularies. This helps maintain consistency in the terminology used for a given data deposit requirement within each repository. The repositories, however, recommended different kinds of controlled vocabulary—even for the same metadata element. Such confusing guidelines make adopting a common language within a discipline unnecessarily difficult.

Compared to the archaeological repositories, fewer repositories in zoology or the social sciences suggest the use of controlled vocabularies. One zoological repository (Morphbank) offers a taxonomic hierarchy for browsing taxon names and a search function for their retrieval. The taxonomic classification is based on the Integrated Taxonomic Information System (ITIS) maintained by the US Department of Agriculture. There is no social science repository that requests the use of controlled vocabularies at the time of deposition.

The Fuzzy Lookup analyses show the extent to which the generalized terms occur for the names and/or descriptions of data deposit requirements. Since the generalized terms are mostly based on DC metadata elements, DANS, which employs DC for the requirements, is the only repository to have matches for most generalized terms. The use of the generalized terms differs across repositories in each discipline. Within the archaeology repositories, Title, Creator, Date, Type, and Location are used by more than half of the repositories (see appendix A). According to appendix B, Date is the only term used by the majority of repositories in zoology. Social science repositories commonly employ Title, Date, Format, Coverage, and File. The results reveal that the generalized terms are not typically used for names and/or descriptions of data deposit requirements in the examined repositories. The analysis also shows the range of metadata labels used to name a like concept. For example, different terminology is used for the requirement of Description since various types of content are requested against Description at both the study and data levels. The generalized term Description is used by three or four repositories in each discipline, which primarily require details on the study background or the data content. One repository in social sciences uses the term "sample description," which relates to the methodologies. A common term across the disciplines for indicating the study background is "abstract." Other terms are also used, including "introduction" (ADS), "purpose," and "research program(s)" (CPN). Moreover, the same term often has multiple meanings across repositories. For example, forms define the Type requirement in a variety of ways. Archaeology repositories (ADS, tDAR, and Parks Canada) require information on the types of media and/or types of materials comprising data. Social sciences repositories (ADA and ICPSR) ask for types based on data collection.

The findings suggest that the metadata requested during data deposition should be defined in a more standardized and consistent manner, particularly those that can have various meanings within and across disciplines. This can help depositors better understand what metadata and documentation is required to prepare for deposition and assist curators to resolve inconsistencies. The use of standardized terminology can also support the interoperability of metadata across repositories, an innovation that would facilitate interdisciplinary data reusers' research.

Need for specific information about the data usage license

The FAIR principles for reusability also advocate for clear and accessible data reuse licenses. This means that the conditions under which data can be reused need to be explicitly stated. As seen from our study, this information can be collected as part of the deposition process. Similarly, the recently published report *Research Data Curation: A Framework for an Institution-wide Services Approach* states, "it is crucial to describe the terms of use, including who can use the data, how the data could be used, and privacy and intellectual property issues," in addition to describing the data.[27] Among the repositories in our study, all social science repositories requested information about either the license agreement or terms of use, as part of the descriptive metadata. Archaeological repositories conceptualized a data usage license as rights information and rights information was requested by most of the archaeological repositories. Social science repositories ask about both data usage licenses as well as rights.

DANS requires information on copyright, license, and/or access rights. With respect to zoology, only CPN requires information on access rights. In particular, CPN provides an option to limit access to data relating to traditional knowledge from the indigenous community of the Polar Regions.[28] The lack of attention to this in zoology is interesting, particularly for the museums represented since repositories often embargo location information for endangered species or ask data reusers not to publish it.[29] Our findings indicate that repositories rarely request the conditions for data reuse; but, when they do, the specific type of license, rights, or access conditions requested require specification.

Conclusion

We investigate both common and unique data deposit requirements and levels of description for each requirement, as identified by 20 repositories in the three disciplines: six in archaeology, eight in zoology, and six in quantitative social science. Based on the qualitative content analyses and Fuzzy Lookup comparisons, we suggest that data deposit requirements relating to Creator, Description, Contributor, Date, Relation, and Location are common across the disciplines. Repositories across disciplines require Publisher and Language the least. Repositories commonly list more requirements in archaeology and quantitative social science than those in zoology. Concerning the levels of description, data-level description requirements are most common, while study-level and/or file-level description are identified far less often, especially by the zoological repositories when compared to the others.

We discuss the implications of the findings based on the FAIR principles, particularly those regarding data reusability and interoperability. Relating to the reusability principle, the repositories must specify data deposit requirements with the study-level description in sufficient detail, since they have requested different types and levels of granularity for contextual information across disciplines. It is also important for the repositories to determine data usage licenses, access conditions, and rights to make data reusable in a legal and ethical manner. In terms of the interoperability principle, the repositories should describe the metadata in a more standardized and coherent manner. This would enable the interoperability of metadata among data repositories, which in turn would support the interdisciplinary research conducted through data reuse.

Acknowledgements

This research was made possible in part by a National Leadership Grant from the Institute of Museum and Library Services, LG-06-10-0140-10, “Dissemination Information Packages for Information Reuse” (DIPIR), with support from OCLC and the University of Michigan School of Information. We thank members of the DIPIR team, including University of Michigan students, research fellows, institutional partners, and individual collaborators. We also thank manuscript reviewers and editors for their insightful comments and suggestions.

APPENDIX A. Fuzzy Lookup Comparison of Data Deposit Requirements against Generalized Terms in Archaeology

Repositories Generalized terms		ADS	tDAR	DANS	Open Context	Parks Canada	Institute for Archaeologists
DC Core Elements	Title	Title (1.0000)	Basic Information: Title (0.8909)	Title (1.0000)	Title (1.0000)		
	Creator	Data Creator (0.9000)	Author/Creator: Person (0.8667)	Creator (1.0000)	Creator (1.0000)		All original written documents created throughout the course of the project (0.7163)
	Subject	Subject (1.0000)		Subject (1.0000)		Image: Subject (0.9429)	
	Description		Basic Information: Description (0.8909)	Description (1.0000)	Short Description (0.9000)	Stratigraphy: Description (0.9000)	
	Publisher		Basic Information: Publisher (0.8909)	Publisher (1.0000)			Other work published during the life of the project (0.7902)
	Contributor			Contributor (1.0000)			
	Date	Project Dates (0.8068)	File Information: Date (0.8833)	Date (1.0000)	Data Format and Structure (0.7621)	Image: Date (0.9200)	
	Type	Data Type (0.0900)	Material Type(s) (0.8615)	Type (1.0000)		Site: Environment: Soil Type (0.8500)	
	Format			Format (1.0000)	Data Format and Structure (0.8526)	Media: Format (0.9250)	
	Identifier	Identifiers (0.9733)	Item-Specific or Agency Identifiers (0.8153)	Identifier (1.0000)			
	Source		Source & Related Comparative Collections: Source Collection (0.8333)	Source (1.0000)		Stratigraphy: Date and Source of Deposit (0.8417)	

Generalized terms \ Repositories		ADS	tDAR	DANS	Open Context	Parks Canada	Institute for Archaeologists
DC Core Elements	Language	Language (1.0000)		Language (1.0000)			
	Relation			Relation (1.0000)	Suboperation: Relationship to Period Features/ Structures (0.7084)		
	Coverage		Temporal Coverage: Temporal Terms (0.8526)	Temporal Coverage (0.8889)			
	Rights	Access Rights (0.9000)		Rights Holder (0.9000)			
Additional Elements (Non-DC)	Comments					Suboperation: Comments (0.9000)	
	File		File Information: File (0.8727)				
	Document						The schedule of works or similar documents (0.7993)
	Location	Location (1.0000)	Basic Information: Publisher Location (0.8625)		Location (1.0000)	Site: Location: Location (0.8800)	

APPENDIX B. Fuzzy Lookup Comparison of Data Deposit Requirements against Generalized Terms in Zoology

Repositories Generalized terms		AMNH	CPN	DANS	Dryad	NCBI GenBank	Morphbank	MVZ	Protein Data Bank
DC Core Elements	Title		Title of the Data (0.8571)	Title (1.0000)					
	Creator	Date of Curator's Signature (0.7357)		Creator (1.0000)					
	Subject			Subject (1.0000)					
	Description	Specimen # or Number of Specimens with Description (0.8299)		Description (1.0000)	Description for each file or group of files: type(s) of data included (0.8158)		Locality: Locality Description (0.8857)		
	Publisher			Publisher (1.0000)	Citation(s) of your published research derived from these data (0.7983)		Image: Date to Publish (0.7919)		
	Contributor			Contributor (1.0000)			Specimen: Contributor (0.9500)		
	Date	Date of Curator's Signature (0.8400)	Data Access (0.8510)	Date (1.0000)	Contact Information for Author(S) Regarding Data Analyses (0.7483)	Release Date Information (0.8727)	Specimen: Date Determined (0.8750)	Date(s) Found (0.8667)	
	Type			Type (1.0000)	Description for Each File or Group of Files: Type(s) of Data Included (0.8211)	Submission Category and Type (0.8500)	Specimen: Relationship Type (0.8750)		

Repositories Generalized terms		AMNH	CPN	DANS	Dryad	NCBI GenBank	Morphbank	MVZ	Protein Data Bank
DC Core Elements	Format			Format (1.0000)			Specimen: Form (0.7150)		
	Identifier			Identifier (1.0000)					
	Source			Source (1.0000)		Organism Name, Applicable Source Modifiers, Location (0.8333)	Taxon: Name Source (0.9000)		
	Language		Plain Language Summary (0.8667)	Language (1.0000)					
	Relation			Relation (1.0000)					
	Coverage			Temporal Coverage (0.8889)					
	Rights			Rights Holder (0.9000)					
Additional Elements (Non-DC)	Comments								
	File				Description for Each File or Group of Files: Type(s) of Data Included (0.8211)				
	Location					Organism Name, Applicable Source Modifiers, Location (0.8333)		MVZ Location (0.9000)	

Repositories Generalized terms		AMNH	CPN	DANS	Dryad	NCBI GenBank	Morphbank	MVZ	Protein Data Bank
Additional Elements (Non-DC)	Sequence					Nucleotide Sequence(s) (0.8545)			Information about the composition of the structure (sequence, chemistry, etc.) (0.8216)
	Taxon						Taxon: Contributor (0.8800)		

APPENDIX C. Fuzzy Lookup Comparison of Data Deposit Requirements against Generalized Terms in Quantitative Social Science

Repositories Generalized terms		ADA	DANS	ICPSR	Roper Center	The Odum institute	UKDA
DC Core Elements	Title	Study Title (0.9250)	Title (1.0000)	Title of the Data Collection (0.8769)	Title of the Survey (0.8667)	Descriptive Title of the Data (0.8653)	Title (1.0000)
	Creator		Creator (1.0000)				Data Creators (0.8807)
	Subject		Subject (1.0000)			Subject Terms (0.9000)	Subject Categories (0.8033)
	Description	Related Publications: Description of Publication (0.8306)	Description (1.0000)	Description or Abstract (0.8714)	Sample Description (0.9143)	Description of Data Contents (0.8667)	
	Publisher		Publisher (1.0000)				
	Contributor		Contributor (1.0000)				
	Date	Date of data Collection (0.8610)	Date (1.0000)	Data Collection Dates (0.8295)	Interview Dates (0.8170)	Date(s) of Data Collection (0.8571)	Data Collectors (0.8084)
	Type		Type (1.0000)	Type of Data (0.9111)	Sample Type (include geographic coverage) (0.8455)	Type of Data (0.9111)	
	Format	Quantitative Data Files: File Description: File Format (0.8533)	Format (1.0000)		Preferred Formats Include ASCII, SPSS, SAS, or STATA (0.7752)	Data Format (0.8667)	
	Identifier		Identifier (1.0000)				
	Source		Source (1.0000)	Type Of Data Collection: Program Source Code (0.8533)		Source(s) of Data (if derived from another data file or from a printed source) (0.8182)	Source Location and Access (0.8526)

Repositories Generalized terms		ADA	DANS	ICPSR	Roper Center	The Odum institute	UKDA
DC Core Elements	Language		Language (1.0000)				
	Relation		Relation (1.0000)				
	Coverage	Geographic Coverage (0.9000)	Temporal Coverage (0.8889)	Geographic Coverage Area(s) (0.8625)	Sample Type (include geographic coverage) (0.8455)	Universe: Geographic Coverage (0.8667)	Geographical Coverage (0.9111)
	Rights		Rights Holder (0.9000)				
Additional Elements (Non-DC)	Comments	Quantitative Data Files: File Description: File Contents (0.7713)				Description of Data Contents (0.7845)	
	Confidentiality			Is there confidential information in the data? (0.7721)			Confidentiality / Anonymization (0.8108)
	Document	Deposited Documentation, Publications and Reports: Description of Document (0.8370)					Observation Units: Text Units (documents / chapters / words) (0.8029)
	File	Deposited Documentation, Publications and Reports: File Name (0.8167)		Select Files to Deposit (0.7653)	Data File(s) (0.8769)	File Layout (if raw format deposited) (0.8294)	Files Being Transferred for Deposit (0.7501)
	Location				Location of the Weights in the Study and a Description of the Weighting Factors (0.8179)	Location (city/ state) of Data Producer (0.8435)	Source Location and Access (0.8526)

Notes

[1] Claire C. Austin et al., "Key Components of Data Publishing: Using Current Best Practices to Develop a Reference Model for Data Publishing," *International Journal on Digital Libraries* 18, no. 2 (June 1, 2017): 77–92, https://doi.org/10.1007/s00799-016-0178-2; Lyubomir Penev et al., "Strategies and Guidelines for Scholarly Publishing of Biodiversity Data," *Research Ideas and Outcomes* 3 (2017): e12431.

2. Ayoung Yoon and Helen Tibbo, "Examination of Data Deposit Practices in Repositories with the OAIS Model," *IASSIST Quarterly* 35, no. 4 (2011).

3. Louise Corti, "Data Collection in Secondary Analysis," in *The SAGE Handbook of Qualitative Data Collection*, ed. Uwe Flick (London, UK: SAGE Publications Ltd, 2018), 164–81, https://doi.org/10.4135/9781526416070.n11; Yoon and Tibbo, "Examination of Data Deposit Practices"; Claire Austin et al., "Guidelines for the Deposit and Preservation of Research Data in Canada," Research Data Canada (2015), available online at https://www.rdc-drc.ca/wp-content/uploads/Guidelines-for-Deposit-of-Research-Data-in-Canada-2015.pdf [accessed 29 July 2019]; UK Data Archive, "Documenting Your Data," available online at https://www.ukdataservice.ac.uk/manage-data/document [accessed 2 July 2018].

4. Ixchel Faniel et al., "The Challenges of Digging Data: A Study of Context in Archaeological Data Reuse," in *Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries*, July 22–27, 2013, Indianapolis, IN (ACM, 2013), 295–304.

5. Ixchel M. Faniel, Adam Kriesberg, and Elizabeth Yakel, "Social Scientists' Satisfaction with Data Reuse," *Journal of the Association for Information Science and Technology* 67, no. 6 (2016):1404–16.

6. Rebecca D. Frank, Elizabeth Yakel, and Ixchel M. Faniel, "Destruction/Reconstruction: Preservation of Archaeological and Zoological Research Data," *Archival Science* 15, no. 2 (2015):141–67.

7. Brian Lavoie, *The Open Archival Information System (OAIS) Reference Model: Introductory Guide*, 2nd ed. (DPC Technology Watch Report 14-02 Oct. 2014).

8. Yoon and Tibbo. "Examination of Data Deposit Practices."

9. Sharon Farnel and Ali Shiri, "Metadata for Research Data: Current Practices and Trends," in *International Conference on Dublin Core and Metadata Applications* (2014), 74–82.

10. Massimiliano Assante et al., "Are Scientific Data Repositories Coping with Research Data Publishing?" *Data Science Journal* 15 (2016), https://doi.org/10.5334/dsj-2016-006.

11. Claire C. Austin et al., "Key Components of Data Publishing: Using Current Best Practices to Develop a Reference Model for Data Publishing," *International Journal on Digital Libraries* 18, no. 2 (2017): 77–92, https://doi.org/10.1007/s00799-016-0178-2.

12. Dimitris Rousidis et al., "Metadata for Big Data: A Preliminary Investigation of Metadata Quality Issues in Research Data Repositories," *Information Services & Use* 34, no. 3/4 (2014): 279–86; Dimitris Rousidis et al., "Evaluation of Metadata in Research Data Repositories: The Case of the DC. Subject Element," in *Research Conference on Metadata and Semantics Research*, 203-13. Springer.

13. Marc, David T., James Beattie, Vitaly Herasevich, Laël Gatewood, and Rui Zhang. 2016. "Assessing Metadata Quality of a Federally Sponsored Health Data Repository." In *AMIA Annual Symposium Proceedings* (American Medical Informatics Association, 2015), 864, available online at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5333273/ [accessed 29 July 2019].

14. Samuelle Carlson and Ben Anderson, "What Are Data? The Many Kinds of Data and Their Implications for Data Re-Use," *Journal of Computer-Mediated Communication* 12, no. 2 (Jan. 2007): 635–51, https://doi.org/10.1111/j.1083-6101.2007.00342.x.

15. For the purposes of clarity during our discussion of metadata types from multiple schema, we format references to DC and non-DC generalized elements by capitalizing their initial letters (such as Title, Subject, and File) and elements identified in data deposit forms by using double quotation marks and preserving original case (like "Site: Environment Soil Type" and "Data format and structures").

16. "DCMI: Dublin Core Metadata Element Set, Version 1.1: Reference Description," available online at http://dublincore.org/documents/dces/ [accessed 13 August 2018].

17. Ibid.

18. Mark D. Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3 (Mar. 15, 2016): 160018, https://doi.org/10.1038/sdata.2016.18.

19. Barend Mons, "FAIR Science for Social Machines: Let's Share Metadata Knowlets in the Internet of FAIR Data and Services," *Data Intelligence* 1, no. 1 (2018), available online at www.data-intelligence-journal.org/p/10/ [accessed 29 July 2019].

20. Swiss National Science Foundation, "Explanation of the FAIR Data Principles," Swiss National Science Foundation, available online at www.snf.ch/SiteCollectionDocuments/FAIR_principles_translation_SNSF_logo.pdf [accessed 17 May 2018].

21. Dutch Techcentre for Life Science, "The FAIR Data Principles Explained," available online at https://www.dtls.nl/fair-data/fair-principles-explained/ [accessed 17 May 2018]..

22. Mons, "FAIR Science for Social Machines."

23. Swiss National Science Foundation, "Explanation of the FAIR Data Principles"; Dutch Techcentre for Life Science, "The FAIR Data Principles Explained."

24. Adam Kriesberg et al., "The Role of Data Reuse in the Apprenticeship Process," *Proceedings of the American Society for Information Science and Technology* 50, no. 1 (Jan. 1, 2013): 1–10, https://doi.org/10.1002/meet.14505001051.

25. Wilkinson et al. "The FAIR Guiding Principles."

26. Craig Willis, Jane Greenberg, and Hollie White, "Analysis and Synthesis of Metadata Goals for Scientific Data," *Journal of the American Society for Information Science and Technology* 63, no. 8 (Aug. 2012): 1505–20, https://doi.org/10.1002/asi.22683.

27. Sayeed Choudhury et al., "Research Data Curation: A Framework for an Institution-wide Services Approach," EDUCAUSE Center for Analysis and Research (May 2018), available online at https://library.educause.edu/resources/2018/5/research-data-curation-a-framework-for-an-institution-wide-services-approach [accessed 29 July 2019].

28. Teresa Scassa and Fraser Taylor, "Legal and Ethical Issues around Incorporating Traditional Knowledge in Polar Data Infrastructures," *Data Science Journal* 16 (2017), https://doi.org/10.5334/dsj-2017-003.

29. Frank, Yakel, and Faniel, "Destruction/Reconstruction."