Scientific Data Repositories on the Web: An Initial Survey

Laura Haak Marcial
School of Information and Library Science, University of North Carolina at Chapel Hill, CB #3360, 205 Manning Hall, Chapel Hill, North Carolina 27599-3360. E-mail: marcial@unc.edu

Bradley M. Hemminger
School of Information and Library Science, 206A Manning Hall, University of North Carolina, Chapel Hill, North Carolina 27599-3360. E-mail: bmh@ils.unc.edu

Science Data Repositories (SDRs) have been recognized as both critical to science, and undergoing a fundamental change. A *websample* study was conducted of 100 SDRs. Information on the websites and from administrators of the SDRs was reviewed to determine salient characteristics of the SDRs, which were used to classify SDRs into groups using a combination of cluster analysis and logistic regression. Characteristics of the SDRs were explored for their role in determining groupings and for their relationship to the success of SDRs. Four of these characteristics were identified as important for further investigation: whether the SDR was supported with grants and contracts, whether support comes from multiple sponsors, what the holding size of the SDR is and whether a preservation policy exists for the SDR. An inferential framework for understanding SDR composition, guided by observations, characteristic collection and refinement and subsequent analysis on elements of group membership, is discussed. The development of SDRs is further examined from a business standpoint, and in comparison to its most similar form, institutional repositories. Because this work identifies important characteristics of SDRs and which characteristics potentially impact the sustainability and success of SDRs, it is expected to be helpful to SDRs.

Introduction

The Internet houses thousands of scientific data centers or repositories (SDRs) in the United States, and it is thought that a much greater number are managed behind firewalls in proprietary environments. Until recently, these sites were primarily developed through government-funded enterprises or within specific domains by groups of self-selected and highly motivated users. Little is known about the universe of SDRs and still less is understood about their lifespan and success, and even what an appropriate definition of success is for an SDR.

This article is focused on gaining an initial understanding of the state of SDRs as seen via the Web. Although scientific collaboration has always involved collections of data, and sharing digital collections occurred before the existence of the Web, the proliferation of Web-based SDRs signals an important change in how scholars utilize these digital collections. Today, anyone can create, contribute data, retrieve data, or annotate existing data in an SDR. The National Science Board (NSB) report (2005, p. 5) note this fundamental change and concluded in its summary: “Long lived digital data collections are powerful catalysts for progress and for democratization of the research and education enterprise.” This easy and convenient access to SDRs provides enormous opportunities, similar to how Web publishing took off in the early days of the Web—intoxicatingly powerful, and easy to get lost in without the benefit of a framework.

The growth of SDRs comes just in time, as the amount of data and increasingly ‘born digital’ data being generated by scientists is exploding. “Big Data” (Nature, 2008) is a 21st century phenomenon. With the recent availability of real-time data collection and enormous strides in computing power and storage capacity, our ability to collect vast amounts of data is burgeoning. In recent months, discussions across scientific domains have focused on how to manage these data and maximize our potential use for it while also minimizing the burden of maintaining it (NSB, 2005; Interagency Working Group on Digital Data [IWGDD], 2009). Efforts like the human genome project demonstrate a newfound capacity to collaborate at a global scale yet these collaborations still remain fairly entrenched in scientific domains. As is noted by

Received December 10, 2009; revised February 16, 2010; accepted February 16, 2010

© 2010 ASIS&T • Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/asi.21339

Borgman, Wallis, and Enyedy (2007), many scientists have begun to embrace the principle of data sharing but the process of exchanging raw data is still maturing.

Emerging SDRs are building on recent ideas like collaboratories (Wulf, 1989), shariums (Marchionini, 1998), and the cyberinfrastructure (David, 2004). They are utilizing practical tools from related domains, including institutional repositories (Eprints, DSpace, & Fedora), digital libraries and publishing worlds, and e-infrastructure (I Rule Oriented Data Systems or iRODS and gCube; Rajasekar, Moore, Wan, & Schroeder, 2009; Aschenbrenner et al., 2008). Attempting to manage larger and larger amounts of information from disparate data types, including everything from molecular scanners to telescope arrays, science data repositories have been described as unique opportunities for scientific scholarly collaboration. The significance and long-term importance of SDRs has also attracted the attention of commercial companies. This is evidenced in announcements like Google's (Madrigal, 2008), that it will create a science data repository on the Web as a "cloud" based service. Google's intent is to also include the capture of so-called "dark data" or data which may never have been refined and/or published. Certainly recent reports from the NSB and the IWGDD suggest that SDRs are a critical element of new science and a national priority.

Because Web-based SDRs are poised to play such a critical role, and because keys to the sustainability and success of SDRs remain elusive, it is important to identify and characterize them. This article attempts to do this, by examining the description of SDRs presented on their Web pages and analyzing the resulting characteristics. Of particular interest is identifying groupings of SDRs that have similar characteristics. A resulting framework inferred from these data could help describe differences among SDRs and help elucidate key elements of sustainability and success. The Web presence of an SDR is the "face" with which the scientist interacts in today's world. This face is the primary mechanism used by the scientist to understand the goals of the SDR, to learn how to interact with it, and to submit and retrieve data. Ideally, to develop a comprehensive picture of SDRs would require understanding each domain field, the scientists, and their information behaviors, as well as organizational issues and the context of the SDR. It is impractical to do this in depth across all domains, so the approach of this study is to take an exploratory look at 100 Web based SDRs from multiple domains to begin to understand them and look for common characteristics.

The goals of this study were to do the following for the sampled SDRs:

- Take an *inventory* of a sizable (100) convenience sample of existing SDRs
- Identify the major *characteristics* of SDRs
- Examine *commonalities* across SDRs
- Look for *trends* over time with respect to SDRs
- Look for characteristics of SDRs that may correlate with the *success* (Maron, 2008) of SDRs

Methods and Results

Inventory

A Web site inventory study was conducted by reviewing possible SDRs. Data collection began in the fall of 2007, followed by an initial evaluation conducted in 2008. Data were sent out to SDR administrators for review and comment in the spring of 2009 and the final evaluation was completed in late 2009. An initial set of SDRs was identified through Google searches, using the following terms: science data center(s), science data repository, scientific data repository, and science digital repository. This was supplemented by investigating the "related links" pages of initial SDRs. The study aim was to identify 100 SDRs, with care taken to try and include repositories of varying disciplines, sizes, and business types. A total of 142 SDRs were identified, of which 100 were included in the final analysis. SDRs that were excluded were extensions of library services, tools, data services, or systems (library, n = 4; tool, n = 2; data services, n = 3; systems, n = 2), those which functioned more like portals (n = 20) and did not contain actual datasets themselves, and SDRs that were moved, replaced, or no longer in existence (n = 11).

Though it is difficult to give a reasonable estimate of the number of SDRs in the current "universe," there are clearly thousands in existence. Many hundreds can be identified from links of SDRs identified here, as well as institutional repository (IR) initiatives (see http://maps.repository66.org/ for details on IRs). The 2005 NSB report (NSB, 2005) suggests that there are hundreds or even thousands of National Science Foundation funded digital collections. In the genetics and biology domain, there are 1,230 identified in the 2010 annual *Nucleic Acids Research* Database Issue (Cochrane & Galperin, 2010). Although there are many differences in the size, types, and organizations of SDRs, there are clearly a large and growing number of them. The SDRs included in this study cover a wide range of uses, including data to accompany published works to data to be used in genetic sequencing. The majority of the SDRs investigated here were examples of highly specific use cases. Some were much broader in terms of both potential use and the heterogeneity of offerings, as in the case of the Odum Institute or of the World Data Center (WDC) for Human Interactions in the Environment.

In general, this study supports the thinking that the earliest of the SDRs were borne of the need to collectively share or store large amounts of data. This occurred initially in particular areas of science. Examples of this include instrumentation that is rare and expensive, in which collaboration occurs around the device and resulting data are shared among research groups (for instance, telescopes, colliders). Another example is when many different disciplines want access to information generated from another area (earth and environmental science, social sciences). Increasingly, sensor data must be aggregated data across a variety of sensors (biomedical informatics, public health informatics, security). In response to the very interdisciplinary nature of modern science, the call to create standard mechanisms for data stewardship and management is being broadcast at the level of funding agencies (IWGDD, 2009).

Characteristics

SDR Web sites were analyzed to identify and describe characteristics of the SDRs. Because this appears to be the first Web-based survey of SDRs, the aim was to identify a wide variety of characteristics of SDRs presented on their Web sites, then examine how these relate to ones described as important in previous literature, and analyze whether groupings of SDRs might exist based on these characteristics and how each characteristic might be contributing to these groupings within the SDR landscape. A complete list of the SDR characteristics is shown in Table 1 below. These characteristics were intentionally defined at a fine grain of detail because grouping or conglomerating characteristics would have made comparisons more difficult and could have unnecessarily introduced bias in deciding how to group them. After the initial identification stage, characteristics were compared with previously established schemes from the literature when possible (column 5 in Table 1), or else described as richly as possible (generally free text). This helped to ensure the collection of more detailed data and it inhibited the degree to which comparisons could be made across the group of 100 SDRs. To facilitate comparisons, a subset of characteristics was refined, reviewed (by SDR administrators), and analyzed to achieve both reliability and homogeneity across the study.

Although data collection from this Web sample is cost effective when considering the Web presence of SDRs, there are several limitations. Obviously, only the information presented on the Web is potentially captured. Furthermore, it is difficult to conclude that a feature is not present because of either a lack of clear evidence or a complete lack of evidence (which can be qualitatively different). To check the validity of the sample data, e-mail was sent to each SDR, providing them the descriptive results for their SDR and requesting clarifications or corrections. Sixty-one SDR administrators provided responses, most of which required no changes (39%, 24 of 61) or minor changes (25%, 15 of 61). Of the total quantitative characteristics analyzed (17 per site, see Table 2) for all the reporting sites, only 1.8% of the 1,037 possible (61 sites each with 17 characteristics) requested changes. The largest number of changes for a single characteristic was 13%. Assuming this rate of corrections holds for the nonresponding sites, this suggests there are no significant problems with the data.

There were changes and additions to the qualitative data collected and several respondents indicated an awareness that their Web site may not present all of this information as fully as needed. Moreover, there was a lot of interest in learning about the survey responses of others and in particular in learning about data collection procedures as well as preservation policies in use.

Of the 50 characteristics recorded for each SDR, a group of 17 (Table 2) were deemed suitable for statistical analysis. The aim of the analysis was to study interactions among the characteristics and to perhaps uncover some similarities of SDRs. Close review of the data indicated a strong association between the HoldingSize variable and the NSB 2005 report's classification of data collections into research, community, or reference categories (on the basis of size, impact, and funding). Another characteristic HowBased was selected to represent the NSB 2005 report categorization of data collections as governmental data centers, university-based consortia, or data federations. In sum, 10 of the characteristics were reasonably represented as binary (NaturalScience, Virtual, InstrumentBased, Centralized/Distributed, SubscriptionorMembership, Multi-Sponsored, GrantsContracts, PreservationPolicy, AcceptSubmittedData, Portal). Another three of the characteristics were suited to an ordinal classification (HoldingSize, RegistrationRequired, and FreeinthePublicDomain). The remaining four variables ScienceArea, BusinessType, Research/Community/Reference, and HowBased were captured as nominal variables. In the analysis, this final set of characteristics (listed in Table 2) appears to be important in defining "group membership" and may eventually be important predictors of SDR sustainability and success (Maron et al., 2009, p. 27).

The first step of the study was to identify key components of SDRs, including the data services offered (type of data, domain, format of data, ingest, export, handling procedures, including archival, storage, curation, preservation, and statistics on use) and business characteristics of SDRs (sponsorship, management, partners, funding vehicles, and governmental affiliation). These data, it was thought, might offer clues to predicting the success of an SDR. Overall, these data were found to be widely heterogeneous across the group of 100 SDRs and were best discussed through descriptive results. The following sections provide qualitative analyses, attempting to characterize the SDR landscape more generally. A spreadsheet containing the complete details of all the characteristics for each SDR is available at http://ils.unc.edu/bmh/pubs/SDR_final_sheet.xlsx.

Scientific domain. SDRs are often described as highly domain specific (Palmer, Cragin, Heidorn, & Smith, 2007); moreover, across domains, they are heterogeneous in their approaches to data sharing and handling procedures. They appear also to be quite different in terms of business characteristics. The SDRs observed in this study were chronicled across a wide variety of domains (see Figure 1). Other characteristics like file types in use, preferred metadata standards and deposit and access details help provide a more complete picture of the relative role of domain in the nature of SDRs.

Research, community, or reference. A characteristic labeled "research/community/reference" was captured as part of this study, referring to rather traditional definitions for these terms and interpreted without attribution to size or funding but more to the actual functionality of the enterprise. For every SDR, the characterization "research" is appropriate on

TABLE 1. Description of characteristics collected in the full data set.

Category	Characteristic	Description	Type	Referenced in. . .
General				

1	Government	Does the SDR appear to be primarily government based?	Y = Yes
			N = No
				Referred to as in-agency and out-agency in 2005 NSB report, p. 24

2	Government/ DataFederation/ UniversityConsortium	Does the SDR appear to be government based, a data federation or a university-based consortium?	G = Government
			DFed = Data Federation
			U = University consortium
				2005 NSB report, p. 15

3	**NaturalScience** or Social sciences	Is the SDR primarily natural or social science focused?	N = natural
			S = social
				2005 NSB report, p. 14

4	**Science area**	Which scientific area applies to the SDR?	Astronomy, biology, chemistry, ecology, geosciences, marine, mathematics, medicine, multidisciplinary, physics, social	Both self-described and generally accepted scientific areas
5	Scientific category	Which science category applies to the SDR?	Astronomy, biology, chemistry, earth, environmental science, hydrology, n/a, physics, planetary/astronomy, social	Initial attempt at classification, some too broad, others too specific

6	**Research/reference/ community**	Which of the following descriptors are most applicable to the SDR?
		In general, responses to this characteristic focused on a typical definition of these terms in which research applied to nearly every SDR and only small subsets could be considered community-centric or reference-like.
			Res = research
			Ref = reference
			Com = community
				Implemented differently than as defined in the 2005 NSB report, p. 14: **See HoldingSize as surrogate for NSB report characteristic**

7	**InstrumentBased**	Is the SDR centered on an instrument or set of instruments?	Y = Yes
			N = No
				2005 NSB report, p. 30

8	**Centralized/Distributed Collection(s)**	Do the SDR's collections appear to be mostly centralized or distributed?	C = Centralized
			D = Distribute
				2005 NSB report, p. 14

9	Presence (**Virtual** or Both)	Does the SDR appear to be borne out of a physical organization or does it appear to be a 'virtual' collection?	V (virtual) or B (both)	Precursor to the centralized/distributed variable, somewhat different in nature
10	**Holding size**	Do the holdings of the SDR appear to be small or specific, moderate in size or scope/breadth or a large, broad holding?	1 = small/less broad, 2 = medium/broad, 3 = large/more broad These descriptions were loosely interpreted but SDR administrator feedback (1 change of 61 responses) showed that these classifications were acceptable.	Most closely related (nearly equivalent) to 2005 NSB report descriptions of Research, Reference and Community digital data collections.
11	Information areas	A brief list of information areas covered by the SDR	Free text with link(s) as appropriate	Background
12	Brief history	A brief history of the SDR	Free text with link(s) as appropriate	Background
13	Description	A brief description of the SDR	Free text with link(s) as appropriate	Background
14	Inception date	Year that SDR appears to have been created	YYYY	Background
15	Contact	Does the SDR provide contact information (either form based or via email)?	Free text with link(s)/e-mail address(es) as appropriate	Background
Business characteristics:				

16	**Business type**	Does the SDR appear to fit one of the business types in the following list:
		non-profit; corporate entity; institute; society; publisher ; university center; federal center; state governmental agency ; partnership; world data center; other
			N-P = Non-Profit
			I = Institute
			S = Society
			Pub = Publisher
			UC = University Center
			FC = Federal Center
			SGA = State Govt. Agency
			P = Partnership
			WDC = World Data Center
			O = Other
				IWGDD, 2009, p. 16

17	**Multisponsored**	Does the SDR appear to have multiple sponsors or a single sponsor?	M = multi, described
			S = single, described
				Maron, 2008, p. 30

18	Sponsorship	What type of sponsorship (see type description) does the SDR appear to have?	S = society
			FGA = federal government agency
			IGA = International government agency
			SGA = state government agency
			OGA = other government agency
			F = foundation
			C = corporation
			I = individual
			U = university
			A = academy
			M = membership
			Sub = subscription
				Maron, 2008, p. 53

19	**GrantsContracts**	Does the SDR appear to obtain primary support from grants and contracts?	Y = Yes
			N = No
				NSB, 2005, p. 23

20	Funding vehicle(s)	Where does the bulk of the funding for the SDR come from (see type list)?	M: <10, <100, <1000, >1000 = Membership
			GD = Government Direct
			S = Subscription (S)
			G = Grants
			C = Contracts
			CA = Cooperative Agreement
			E = Endowment Income
			Svcs = Services
			D = Donations
			ID = Institute Direct
			UNK = Unknown
				Maron, 2008, p. 33

21	Funding	What are the funding details?	Free text with link(s) as appropriate	Background

22	**Subscription Membership**	Does SDR appear to obtain primary support from subscriptions or memberships?	Y = Yes
			N = No
				Maron, 2008, p. 33

23	Subscription/ membership details	What are the details of the subscription/membership model?	Free text with link(s) as appropriate	Background

24	**How based**	Does the SDR appear to be an independent entity, based in a university, completely government based, or an aggregate of these?	Independent (I);
			University (U);
			Government (G);
			Aggregate (A: describe)
				Most closely related (nearly equivalent) to 2005 NSB report, p. 15 description of data collections as government data centers, university consortia or data federations. Data federations can be broken down into independent or aggregate.

25	Structure within organization	What appears to be the structure of the SDR within its organization?	Free text with link(s) as appropriate	Background
26	Management	What is indicated in terms of the management of the SDR?	Free text with link(s) as appropriate	Background
27	Partners	Who are listed as partners to the SDR? This is differentiated from members or subscribers.	Free text with link(s) as appropriate	Background
Data details/policies:				

28	**AcceptSubmittedData**	Does the SDR accept submitted data?	Y = Yes
			N = No
				IWGDD, 2009, p. 19

29	Ingest process	If the SDR accepts submitted data, what is the ingest process?	Free text with link(s) as appropriate	IWGDD, 2009, p. 19
30	Submission details	What are the details of the submission process?	Free text with link(s) as appropriate	Background

(*Continued*)

TABLE 1. (*Continued*)

Category	Characteristic	Description	Type	Referenced in. . .
31	File formats accepted	If the SDR accepts submitted data, what file formats does the SDR accept on submission?	See file format details	IWGDD, 2009, p. 14
32	Submission methods supported	If the SDR accepts submitted data, what methods of submission are supported?	Free text with common methods aggregated and link(s) as appropriate	Background
33	Copyright details	If the SDR accepts submitted data, what are the import copyright details?	Free text with link(s) as appropriate	NSB, 2005, p. 26
34	Export data	Does the SDR export data?	All are Y (set as a core criteria for inclusion)	IWGDD, 2009, p. 19
35	Data file formats transmitted	What file formats are data transmitted in?	See file format details	IWGDD, 2009, p. 14
36	Export methods supported	What export methods are supported?	Free text with common methods aggregated and link(s) as appropriate	Background
37	Export data rights	What are the data rights for exported data?	Free text with link(s) as appropriate	NSB, 2005, p. 26

38	**FreeinthePublic Domain**	Is any of the SDR data available for export for free (free in the public domain)?	Y = Yes
			N = No
			D = Depends (typically different for different activities on the site)
				NSB, 2005, p. 18; IWGDD, 2009, p. 15

39	Fees	Are any fees levied for data access?	Free text with link(s) as appropriate	Background

40	**RegistrationRequired**	Does the SDR require user registration in order to access data?	Y = Yes
			N = No
			D = Depends (typically different for different activities on the site)
				Maron, 2008, p. 46
				As SDRs mature, there is a general recognition that usage data are vital to sustainability. To do this, requiring a registration enables the capture of a user profile and/or the triggering of a log of user behavior. This is a critical element in beginning to understand how users find and access the resources of the SDR.

41	Export copyright details	What are the copyright details for exported data?	Free text with link(s) as appropriate	Background
42	Restrictions on use of data	What, if any, are the restrictions on use of the exported data?	Free text with link(s) as appropriate	Background
43	Attribution statement	Is an attribution statement outlined when exported data are used?	Free text with link(s) as appropriate	Background

44	**PreservationPolicy**	Does the SDR have a preservation policy?	Y = Yes
			N = No
				IWGDD, 2009, p. 15

45	Preservation policy details	Is a preservation policy outlined?	Free text with link(s) as appropriate	Background
Other services:				
46	Data services	Does the SDR provide extensive or additional data services?	Free text with link(s) as appropriate	Maron, Smith, & Loy, 2009, p. 24
47	Data access details	What are the details for more involved data access or other kinds of data access?	Free text with link(s) as appropriate	Background
48	Other services	Are there other services (publications, education, etc.) provided by the SDR?	Free text with link(s) as appropriate	Maron et al., 2009, p. 24

49	**Portal**	Is the SDR also functioning as a portal to additional data repositories?	Y = Yes
			N = No
				Maron et al., 2008, p. 29

50	Data collection on use	Does the SDR capture and make available data on use of the repository?	Free text with link(s) as appropriate	Maron et al., 2009, p. 27

TABLE 2. Description of the 17 characteristics derived from the full data set selected for data analysis.

Category	Characteristic	Type
General		
1	NaturalScience	Binary
2	ScienceArea	Nominal
3	Virtual	Binary
4	HoldingSize	Ordinal
5	Research/Community/Reference	Nominal
6	InstrumentBased	Binary
7	Centralized/Distributed	Binary
Business characteristics		
8	BusinessType	Nominal
9	SubscriptionorMembership	Binary
10	HowBased	Nominal
11	Multisponsored	Binary
12	GrantsContracts	Binary
Data details/policies		
13	AcceptSubmittedData	Binary
14	RegistrationRequired	Ordinal
15	FreeinthePublicDomain	Ordinal
16	PreservationPolicy	Binary
17	Portal	Binary

some level, this resulted in only a small subset of SDRs being characterized as being more "community-centric" or "reference-like" than simply research focused. Note that this is different from how some others have used these same terms (NSB 2005); these differences are covered in the Discussion section.

Holding size. Each SDR was categorized by holding size according to the following definitions: 1 = small/less broad, 2 = medium/broad, 3 = large/more broad. Although holding size may apply differently to different aspects of the SDR or its holdings, capturing this as a single characteristic facilitated making gross comparisons across the many disparate groups. The final distribution among the sampled SDRs is shown in Table 3.

The convenience sampling method certainly had an impact on the imbalance in these numbers, though an effort was made to include SDRs in the one-level or two-level category. Although it is unclear whether or not this information is generalizable to the universe of SDRs, this metric is both important and hard to get right as an outside observer to an SDR. Although it clearly has meaning, it became evident in the process of obtaining feedback from SDR administrators that it could be both more informative and more aptly described if broken out into several subcategories, including size of scientific community, impact of holdings, magnitude of holdings, uniqueness of holdings, etc.

Governmentally based SDRs. At the outset of data collection, it was apparent that SDRs, which are directly or mostly funded and closely affiliated with governmental agencies, centers and/or projects, would represent an important group.

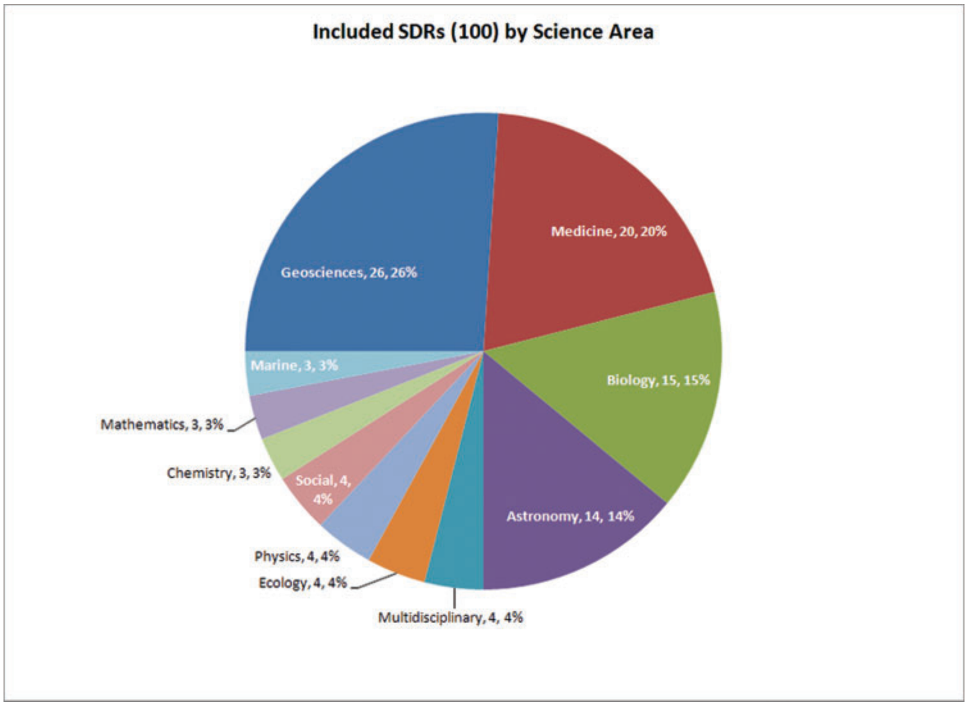


FIG. 1. Distribution of SDRs included in the study by scientific area. Numbers are counts, but also percentages given $N = 100$. These categories have been applied in a mutually exclusive fashion though many of these SDRs have holdings that are representative of multiple scientific areas.

TABLE 3. Final breakdown of Holding Size characteristics, derived among 100 SDRs sampled.

Holding Size	No. SDRs
1 = small/less broad	16
2 = medium/broad	24
3 = large/more broad	60

This distinction was initially captured in detailed information on funding sources, how the SDR was based within an organization and business type. Three characteristics identified in this study—Government, GovernmentDataCenter/DataFederation/UniversityConsortium and HowBased—were used to describe governmental involvement in SDR composition. Each portends related, probably graduated, levels of differentiation. Government was employed as a simple binary characteristic at a high level. The GovernmentDataCenter/DataFederation/UniversityConsortium characteristic further broke down the "nongovernmental" entities into two constituent categories: data federation and university consortia. This was generally in accordance with the NSB 2005 report's characterization along these lines. The HowBased characteristic further differentiated the nongovernmental SDRs with the addition of classifiers "independent" and "aggregate." These tiers were intended to help capture the complex nature of some SDRs. For example, the National Center of Ecological Analysis and Synthesis or NCEAS (http://www.nceas.ucsb.edu/) received its initial and ongoing funding from NSF and is based within a university but functions like a data federation, as evidenced by its collaborations. Conversely, the Amphibian Ark Project (http://www.amphibianark.org/) appears to be supported by a variety of nongovernmental funding agencies, to be functioning as an independent entity with a similar data federation style. Because of the similarities of these variables, only one, HowBased, was used in the main analysis.

Deposits and access. Different aspects of the processes by which data are made available or are accepted were investigated, from data access and deposition policies to submission methods and file types accepted. Data submission/deposition policies vary considerably among the SDRs. Some have detailed guidelines regarding data preparation and Web-based tools for submission. Others offer e-mail contact information and sometimes telephone contact information as a first or primary point of contact. Much of this might be attributed to the degree of heterogeneity of domain, size, and primary purpose of the SDRs included here. In some cases, as in the WDCs and the genetic sequences databases, deposition is a requirement of publication or funding. In several cases, data submissions do appear to be accepted but are nonstandard, so little information, if any, exists to guide the potential end user who is interested in depositing material. In cases where membership is required, it was often not possible to gather data on the actual policies or procedures for deposition. In keeping with the emergence of citizen-based science, there was evidence of highly structured, forms-based deposition of information available for the general public to make submissions.

Provisions concerning data preparation were typically found in relation to submission details rather than general policy guidelines. In many cases, domain specific data preparation schemes such as gene ontology (GO) annotations (Ashburner et al., 2000) are used. In addition, specific exchange protocols are supported to facilitate information transfer and management. An example of this is the acceptance of distributed annotation system (DAS) data—biological annotation data in the DAS communication protocol. At a low level, there appears to be a fairly wide use of an overwhelming number of controlled vocabularies and ontologies to describe data elements. At the next level, there was some adherence to metadata standards for the exchange of information. Finally, whether captured as metadata or as disconnected pieces, the submission would usually be in a particular, well-known file format.

Information representation. File types in use varied widely for ingest and export in the SDRs observed. They range from simple ASCII text to highly specialized formats like Flexible Image Transport System or FITS, a protocol in wide use among astronomers with readable metadata. The relationship between a collection and the file types in use appears to be complex.

- *Ingest*. In the majority of SDRs observed, significant infrastructure exists to support the ingest process. Many domain specific file formats are supported for ingest. Most SDRs indicate support of a wide variety of file types for submission, provide data preparation services, or make extensive documentation available to guide the depositor. In addition, for frequent depositors, much effort has gone into streamlining the deposition process while maintaining quality. It is clear that for the majority of SDRs, which support ingest, encouraging contributions is a key element of their "business practices" and being accommodating is the driver.
- *Processing*. In some cases, sophisticated systems have been built to support the process of data transfer, annotation, or visualization, examples of those include the *Osprey* Network Visualization System used by BioGRID (http://www.thebiogrid.org/) and the bioinformatics community to produce data rich graphical representations using GO annotated data. The BioSystematic Database of World Diptera or BDWD (http://www.sel.barc.usda.gov/Diptera/biosys.htm) comprises a Nomenclator to check names and find basic information for all names and a species database used to answer queries about the attributes of species, such as distribution, biological associates, and economic importance. MapServer (http://mapserver.org/) is an open source platform for publishing spatial data that is used by the Woods Hole Oceanographic Institute Data Center. Morpho, developed by the Knowledge Network for Biocomplexity or KNB (http://knb.ecoinformatics.org/morphoportal.jsp), is a data management software tool used by ecologists, which enables the creation, editing, search, and querying of metadata and the ability to view, edit, and share data (via the KNB), along with an access control layer. To enhance Geographic Information Systems information,

the Socioeconomic Data and Applications Center or SEDAC makes available the SEDAC map client mapping tool (http://sedac.ciesin.columbia.edu/mapviewer/index.jsp), compatible with Open Geospatial Consortium (OGC) standards, which allows for interoperable exchange of map information via e-mail and other services.

- *Export*. For many SDRs, making some data accessible via Web-based tables or images is important. With the advent of registration requirements and membership business models, it is hard in some cases to observe the exact format of the data offerings. In cases of broad, often multidisciplinary SDRs, data are available in a wide variety of formats. Frequently, SDRs make an effort to make data available in convenient formats (html tables, ASCII text flat files, and comma delimited files). There are also a wide variety of image and multimedia formats in use. Many discipline specific file types were observed like FITS for astronomical data, GO/FASTA/Contig annotations for bioinformatics data, statistical formats (SAS, SPSS, R, Stata) for social sciences data, and GIS formats for earth and some biological sciences data. Extensible file formats (XML/EML) tended to be in use among some of the newer SDRs, particularly those focusing on information synthesis (Knowledge Network for Biocomplexity, National Center for Ecological Analysis and Synthesis and Encyclopedia of Life). Table 4 contains a list of file types that were commonly observed on SDR Web sites. Although a large number of different types are utilized, there appears to be a "long tail" effect with a clear preference to reuse common everyday file types, particularly for data export.

TABLE 4. File types commonly observed among the 100 SDRs sampled, particularly for export purposes.

File type category	File type/extension
Archives	.zip, .tar, .tar.gz, stuffit (binhex)
Statistical analysis	R, SPSS, SAS, STATA
GIS	many SDRs indicated using GIS related files including raster formats like .bil, ESRI map file formats like .e00, and vector formats like .shp
Extensible markup	.xml, .sgl, .eml (ecological metadata language), VOTable (Virtual Observatory Table)
Flat file	.txt, .ascii, .csv
Image	.tiff, .jpg, .gif, .pic, .fits and .png
Movie/multimedia	.wav, .swf, .mpg, .mov, .mp3, .mp4, .avi, quicktime and anis (Flash animations applet)
Word processor	.pdf, .ps, .doc
Spreadsheet	.xls
Presentation	.ppt
Proprietary or specific tools:	
Geosciences	Open Geospatial Consortium's Web Map Service (WMS) map and legend images, Web Feature Service (WFS) vector source data in GML format, Web Coverage Service (WCS) raster source data in GeoTIFF format NetCDF (common data format, http://www.unidata.ucar.edu/software/netcdf/docs/faq.html) and .grib (gridded binary)
(Medicine) bioinformatics	GO, FASTA, Contig
Web page	.html

TABLE 5. Ingest method details, as observed among 64 SDRs sampled, which support ingest (not mutually exclusive because many SDRs offer a variety of methods).

Form of data transmission	No. of SDRs
Web-based form or software (including Web services)	26
E-mail	21
FTP	16
Hardcopy	11

Ingest methods. Methods used to accept submitted data vary widely across this group of SDRs but a few methods appear to be particularly popular: Web sites using a Web-based form, ftp/scp, and e-mail. In addition to these, there are many specialized tools and software applications that facilitate data deposition. Some SDRs still prefer to be contacted directly before anything is submitted. In these cases, it isn't clear whether the preference for direct contact initially is a result of low volume use or domain-specific details. There also could be editorial management or data preparation concerns. Although, in the past, it was common for sites to accept submitted data on media like tapes or CDs, the primary method of submission for recently emerging SDRs is to accept data directly over the Internet.

Of the 64 SDRs included in our sample that appeared to support both ingest and export of data, many (35) indicated use of a variety of submission/transport methods. These ranged from sophisticated online forms or software to hardcopies like tapes, flash drives, and CDs. Table 5 shows the range of data transmission routes supported with a rough indication of their relative use:

These data suggest that the sites are, in general, trending toward online forms of submission. Although roughly 29 of these SDRs indicated supporting only one of these methods of data transmission, this is more likely to be the result of how information is presented on the Web sites of these organizations than of a desire to limit modalities of transfer. However, limiting the transfer method may be leading to improved data quality and adherence to standards in some cases. There is indication of a (perhaps) natural relationship between age of SDR and the types and variety of transmission methods supported. Older SDRs tend to be more exhaustive in the transmission options allowed.

Metadata. In general, the presentation of the use of metadata standards by the SDRs in our sample was inconsistent at best. This is likely because of the fact that most SDRs, while they may employ a standard, do not necessarily make this information transparent on their Web sites. Although there was clear indication of several standards in use, it was not apparent that metadata standards have been implemented universally and it could be concluded that metadata standards use, in general, is underreported by this sampling and data collection method. However, although metadata incorporation is not present in all cases and the standards used vary a great deal by discipline, it is clearly an increasing priority.

DOI: 10.1002/asi

What was observed in many cases where Web-based forms are part of the ingest process was Webform-based metadata capture of a descriptive nature. Broadly, the focus remained on descriptive metadata, though system level and policy state metadata may also be in use. It can also be assumed that there are significant differences in metadata use based on data type (experimental, observational, simulation, or derived data product). Though this was not readily observed for this study, it would be an important component of future work. Although this allows for verification of adherence to requested data formats for individual data elements, until common metadata standards are utilized, interoperability will be limited and researchers will not be able to pull data from multiple SDRs for combined analyses, stifling opportunities for Web 2.0 style query and retrieval (Chan & Zeng, 2006).

Usage statistics. Evidence that the SDRs maintained data on submissions, access and use of their resources was reviewed. This information would help elicit characteristics that may affect longevity (Maron et al., 2009). Again, these results were highly variable, but it was important to see that some SDRs were making a clear effort to expose this kind of information to their users. Some SDRs also appear to be using these data to help generate revenue or establish different kinds of relationships with different user constituencies. Though the majority of sites reviewed do not appear to maintain data on contribution and use, some do, and it appears that this information may be an important indicator of success (Maron et al., 2009). For those that do, the examples range from simple disclaimers describing a data collection methodology from a privacy standpoint, as in the case of the CDC (http://www.cdc.gov/doc.do?id=0900f3ec80093c90) to a graphical display of aggregate data on contribution(s) and use (http://www.ddbj.nig.ac.jp/statistics-e.html), and discussion in an already standardized reporting system such as the WDC system (http://www.ngdc.noaa.gov/wdc/reports.shtml). A particularly good example of usage data can be found under the heading "DDBJ Data Submission Activities" on the DNA Data Bank of Japan or DDBJ Web site at http://www.ddbj.nig.ac.jp/documents-e.html. Here, you can follow links to both submissions (by year and by agency) and to server statistics and archival information on use. Many sites do not make their usage data available to the public but maintain it internally and may make it available to members, sometimes in exchange for revenue (Maron et al., 2009, p. 27). In response to our inquiry, several SDR administrators who do not currently collect usage data expressed interest in obtaining examples from those who do.

Business type. A primary goal of the study was to attempt to characterize SDRs in terms of some basic business type characteristics. Where possible, data was collected on funding mechanisms (grants, contracts, gifts, etc.), recording whether an SDR appeared to be primarily government-based/funded with single sponsorship or whether the entity appeared to be more independent or university based with multiple forms of sponsorship. Information on noted partners, organizational details such as structure within an organization, and any basic management details that could clearly be identified were also collected. These factors were used to define classes of "business types" for the purpose of making comparisons. Figure 2 represents a mutually exclusive categorization of the SDRs in our study by Business Type.

Memberships or subscriptions. A small number (n = 16) of SDRs actively supports a substantial fraction of their activities through memberships or subscriptions or requires access through verified membership. In some of these cases, additional support is also provided through governmental entities and grants. Models for memberships or subscriptions varied from that of traditional print publishing, as in the case of the Ecological Society of America (ESA) and institutional membership models like that of the Inter-University Consortium for Political and Social Research (ICPSR) to a more modern electronic data model, as exemplified by the Global Biodiversity Information Facility (GBIF). The GBIF model supports a formalized "data sharing agreement" for its providers, which helps to standardize procedures and expectations for both the participant and the SDR. (IWGDD, 2009, p. 16)

Preservation. Detailed searches were used to gather information on any specific mention of a preservation policy. This was not limited to sustainability, trustworthiness, or interpretability; rather, it was viewed broadly as any indication, discussion, or plans related to the long-term management of data. To be both more exhaustive and more consistent, these searches were broadened from within Web site-based inquiries to Google-based searches. A clear mention of a preservation policy or similar was recorded for 62 (62%) of the SDRs included in the study, with the remaining 38 (38%) either making no mention or no clear mention of such a policy. Policy details varied considerably across the total study group, and a number of groups that share similar governance were assumed to be operating under the same basic preservation policy guidelines (unless otherwise specified). In some cases, a clear indication of curation or archive activities was given without any specific details regarding a preservation policy. This particular characteristic was of definite interest to SDR administrators, who currently do not have a formal policy and examples of policies in current use were forwarded for their information.

Additional services provided by SDRs. In addition to the wide array of services offered by SDRs to manage incoming and outgoing data, many of these SDRs are either borne from organizations that may have originated by offering other types of services or have added additional services to their offerings to meet the needs of their users/depositors. These services include educational offerings, technical assistance including data management and manipulation services (access to computing facilities, curation, archive and preservation tools, and information), print and publication services, marketing, publicity, and software development services.

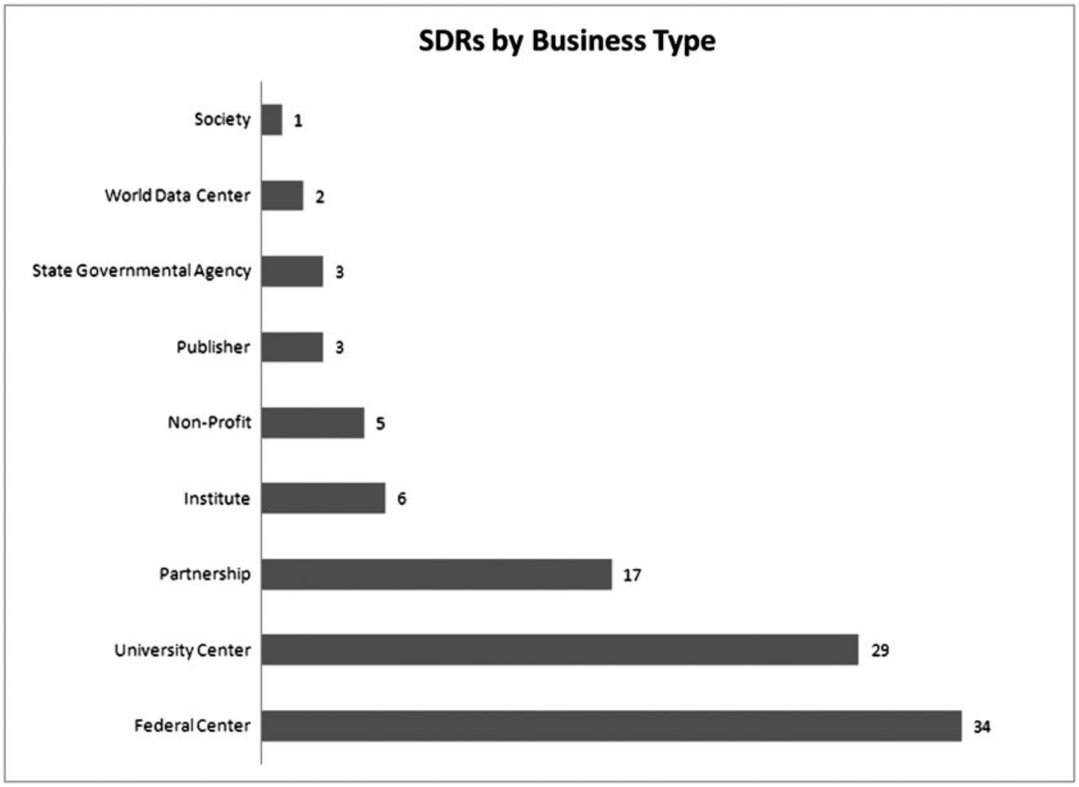


FIG. 2. SDRs by business type. Note that the majority of SDRs included in this study appeared to have direct governmental funding earning their classification as a federal center. The second most prevalent category was university center. Often, funding for these two types comes largely from the federal sector.

They also offer opportunities to collaborate through working groups, task groups, and simulation environments and tools. It may be increasingly important to provide such services as a way to differentiate offerings from other SDRs. This can increase traction with users and help diversify funding opportunities (Maron et al., 2009, p. 24).

Analysis

Identifying and quantitatively coding 17 of the key characteristics of the SDRs allowed for reasonable quantitative analysis aimed at preclassification of the data. Relationships between characteristics (called variables in the analysis section) can be better described through the combined use of cluster analysis and logistic regression used as a post clustering explicatory tool. Being able to identify groups through clusters of similar SDRs facilitates studying the effect of these group characteristics with respect to group membership, sustainability, success, and future trends. This may also enable the development of common data management and stewardship plans or tools and help provide avenues for social networking across domains.

Grouping similar SDRs. Cluster analysis was chosen as a good foundation for an exploratory analysis, using the categorical data collected. This analysis enabled the maximizing of dissimilarity between groups to uncover possible similarities among SDR groups to see if they naturally partitioned in ways that could be explained well with the variables in our model. If a few well-defined classes of SDRs emerged from the cluster analysis, and it was possible to measure how successful they were, then this information could be used to provide models for the successful creation and management of SDRs. If instead many different types emerged with little similarity, then it would be difficult to develop such guidelines or generalizations. Cluster analysis was performed using Ward's method (Romesburg, 2004) on the 17 multinomial variables using PROC DISTANCE to obtain distance measures, PROC CLUSTER to perform the clustering, and PROC TREE to obtain a dendrogram (Johnson, 1967) used to help visualize the clustering results (SAS Version 8.1). The dendrogram in Figure 3 visually depicts the clustering results, showing individual independent SDRs (bottom) as they form into larger and larger groupings (top). As they grow upwards, the smaller clusters merge into larger clusters of SDRs with similar characteristics. A cluster that remains together for a long time (represented by a long line on the vertical axis) demonstrates a persistent set of similar characteristics. The y-axis value is a semipartial r-squared, which gives a measure of degree of differentiation between the groups. Note how Clusters B, C, and D in the four-cluster solution join to form Cluster 2 in the two-cluster solution shown as Cluster 1 and Cluster 2 on the dendrogram.

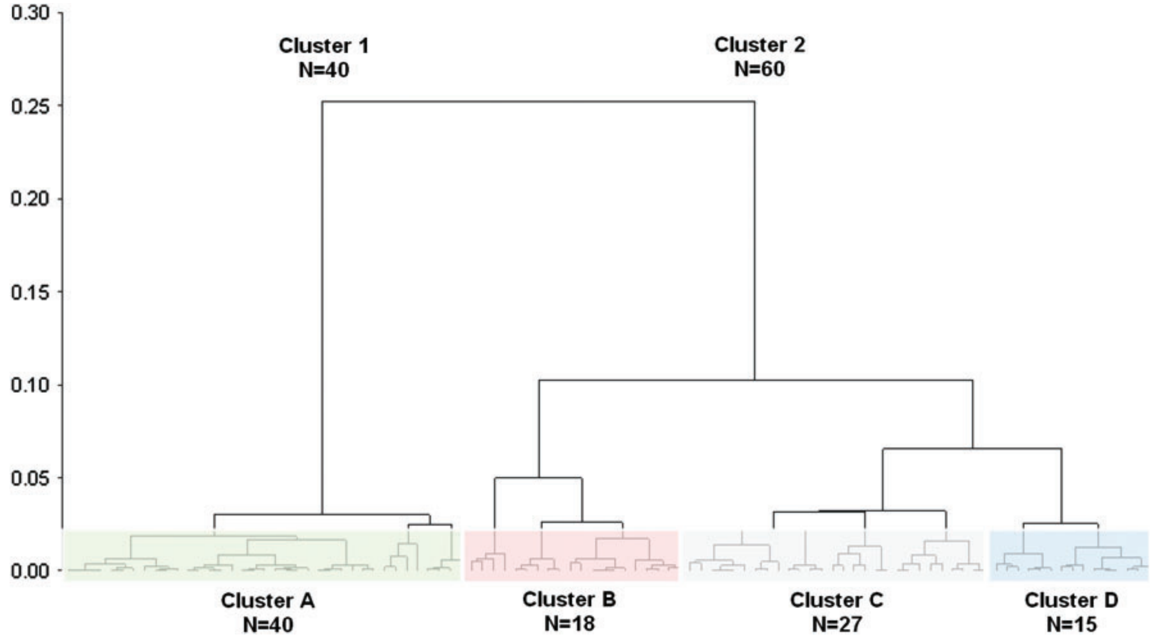


FIG. 3. Dendrogram study comparing cluster formation results. The y-axis shows semipartial r-squared values, a measure of cluster differentiation. The X-axis shows the cluster groupings.

For further analysis, the four-cluster solution, depicted here as clusters A, B, C, and D, was selected for its persistence (as indicated by the semipartial r-squared value on the y-axis) and for the resulting degree of differentiation. This number of clusters also allows for a more detailed exploration of group composition. At this level of clustering, the four groups have been shaded the same in Figure 3 and Table 6, for easy comparison.

To understand more about the composition of the four clusters, cluster membership was incorporated into the dataset and a simple logistic regression was performed (PROC LOGISTIC, SAS Version 8.1) on each variable. Comparison of the Wald chi-square test statistic, divided by the degrees of freedom for each variable (depicted in Figure 4), yields a measure of relative strength of association on cluster membership for each variable when each of the variables is taken independently. As an exploratory analysis, this simple test helps to describe the role each of the characteristics plays in distinguishing between the four clusters (A, B, C, and D) in Figure 3.

Figure 4 shows the variables most responsible for group differentiation: GrantsContracts, MultipleSponsors, HoldingSize, PreservationPolicy, VirtuallyBased, RegistrationRequired, HowBased, AcceptSubmittedData, Centralized/Distributed, and InstrumentBased. Of somewhat lesser strength of association were Research/Community/Reference, Portal, NaturalScience, SubscriptionMembership, ScientificArea, BusinessType, and FreeinthePublicDomain.

The variables not used in the main analysis may also play a role in differentiation. The effects of the remaining variables, though, are harder to standardize and measure. In addition, it can be assumed that some of the effect of these variables has been inherently represented in the related variables included in the analysis.

To go beyond understanding individual variable contribution (Figure 4) and group membership (Table 6), group composition is examined. We performed a simple decomposition by identifying the majority values for each variable in each group (Table 7). Differences in majority values across the groups that were considered qualitatively meaningful are highlighted. From this analysis the relative "group titles" are obtained.

Based on the information in Table 7, it is clear that some variables may be highly correlated, as in the cases of GrantsContracts, MultipleSponsors, AcceptSubmitted-Data, and InstrumentBased. Certainly for establishing the characteristics in Cluster A, the "Governmental" cluster, this is the case. The variables GrantsContracts and MultipleSponsors also play roles in differentiating among the remaining clusters B, C, and D. There is a subset of variables—include NaturalScience, SubscriptionMembership, and FreeinthePublicDomain—that do not play a role at all in the final clustering results, as the values remain consistent across all four groups. This is probably because the number of either yes or no values was too small to generate much of an effect. It is important to note that in the case of the nominal variables with many response options, HowBased (4), Research/Community/Reference (3), ScientificArea (11) and BusinessType (10), seeing any single response option predominate in a group is indicative of the composition of the group and suggests they are particularly noteworthy components of group composition.

The clustering results depend on the variables included. Because the 17 variables used in this analysis were chosen

TABLE 6. The listing of individual SDRs in Clusters A, B, C, and D.

Group A	**Group C**
Agency for Healthcare Quality and Research	ACE Science Center (ASC)
Alternative Fuels Data Center (AFDC)	Antarctic Glaciological Data Center (AGDC)
Atlantic Oceanographic and Meteorological Laboratory (AOML) Environmental Data Server or ENVIDS	Astronomy Digital Image Library
Atmospheric Radiation Monitoring (ARM) Data Centers	Brain biodiversity bank at Michigan State University
Carbon Dioxide Information Analysis Center (CDIAC)	Bugwood Network
Centers for Disease Control and Prevention Data and Statistics	Center for International Earth Science Information Network (CIESIN)
Climate and Environmental Retrieval and Archive (CERA) for the WDCC	Chesapeake Bay Environmental Observatory (CBEO) Portal
Chandra data archive	Coastal Data Information Program (CDIP) of the Scripps Institution of Oceanography, University of California at San Diego
Comprehensive Epidemiological Data Resource (CEDR)	Cornell University Geospatial Information Repository
Controlled Fusion Atomic Data Center (CFADC)	Forestry Images
DNA Data Bank of Japan (DDBJ)	Henry A. Murray Research Archive (MRA)
DOE Joint Genome Instituteís (JGI) Genome Web Portal	IAU Minor Planet Center
DOE's Energy Information Administration (EIA)	Inter-university Consortium for Political and Social Research (ICPSR)
European Southern Observatory (ESO) Archive Facility	IQSS Dataverse network
Genbank	LTER Network
Geodata.gov	McIDAS
NASA's High Energy Astrophysics Science Archive Research Center (HEASARC)	Melanoma Molecular Map Project
HubbleSite Gallery	Repository for Archiving, Managing and Accessing Diverse Data (RAMADDA)
NOAA's Integrated Coral Observing Network (ICON)	Socioeconomic Data and Applications Center (SEDAC)
Integrated Monitoring Network	Space Science and Engineering Center (SSEC) Data Center, University of Wisconsin-Madison
Multimission Archive at STScI (MAST)	The Howard W. Odum Institute for Research in Social Science
NASA Langley Atmospheric Science Data Center	The USA National Phenology Network (USA-NPN)
NASA/IPAC Infrared Science Archive (IRSA)	Thematic Realtime Environmental Distributed Data Services (THREDDS) Data Server
National Ecological Observatory Network (NEON)	Unidata Program at the University Corporation for Atmospheric Research (UCAR)
National Nuclear Data Center Nuclear Data Portal	University of California Santa Cruz Genome Bioinformatics
National Space Science Data Center	Woods Hole Oceanographic Institute Data C enter
Natural Resource and GIS Metadata and Data Store of the National Park Service	World Data Center for Human Interactions in the Environment
Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)	**Group D**
Planetary Data System (PDS)	Amphibian Ark Team Portal
Renewable Resource Data Center (RReDC)	Discover Life in America's Great Smoky Mountains National Park's All Taxa Biodiversity Inventory
Solar Data Analysis Center (SDAC) at NASA Goddard Space Flight Center	Encyclopedia of Life
SkyView	fMRI Data Center
Smithsonian Tropical Research Institute's (STRI) Center for Tropical Forest Science (CTFS)	Global Biodiversity Information Facility
U.S. Transuranium and Uranium Registries (USTUR)	Knowledge Network for Biocomplexity (KNB)
United States Census Bureau	Mouse Genome Informatics
US National Virtual Observatory (NVO)	NEEScentral
US Transplant—Scientific Registry of Transplant Recipients	Netlib
Visible Human Project®	Ocean Biogeographic Information System (OBIS)
World Data Center (WDC)	Paleobiology Database
World Data Center (WDC) for Biodiversity and Ecology	PANGAEA® - Publishing Network for Geoscientific and Environmental Data
Group B	Tree of Life Web Project
BioSystematic Database of World Diptera (BDWD)	Treebase, Treeb ase2
CalSurv, the California Vectorborne Disease Surveillance System	Veg Bank, a vegetation plot database
Ecological Society of America's Ecological Archives	
European Molecular Biology Laboratory- European Bioinformatics Institute or EMBL-EBI	
Encyclopedia of Astronomy and Astrophysics	
Ensembl	
International Council for Science : Committee on Data for Science and Technology	
Iubio	
J. Craig Venter Institute	
Jaspar	
Journal of Applied Econometrics (JAE) Data Archive	
National Center for Ecological Analysis and Synthesis (NCEAS) Data Repository	
NC One Map	
Spec Patterns	
The BioGRID	
The Sanger Institute	

Note. The same highlight coloring that is used here and in Figure 3 shows membership in the cluster groupings depicted in Figure 3.

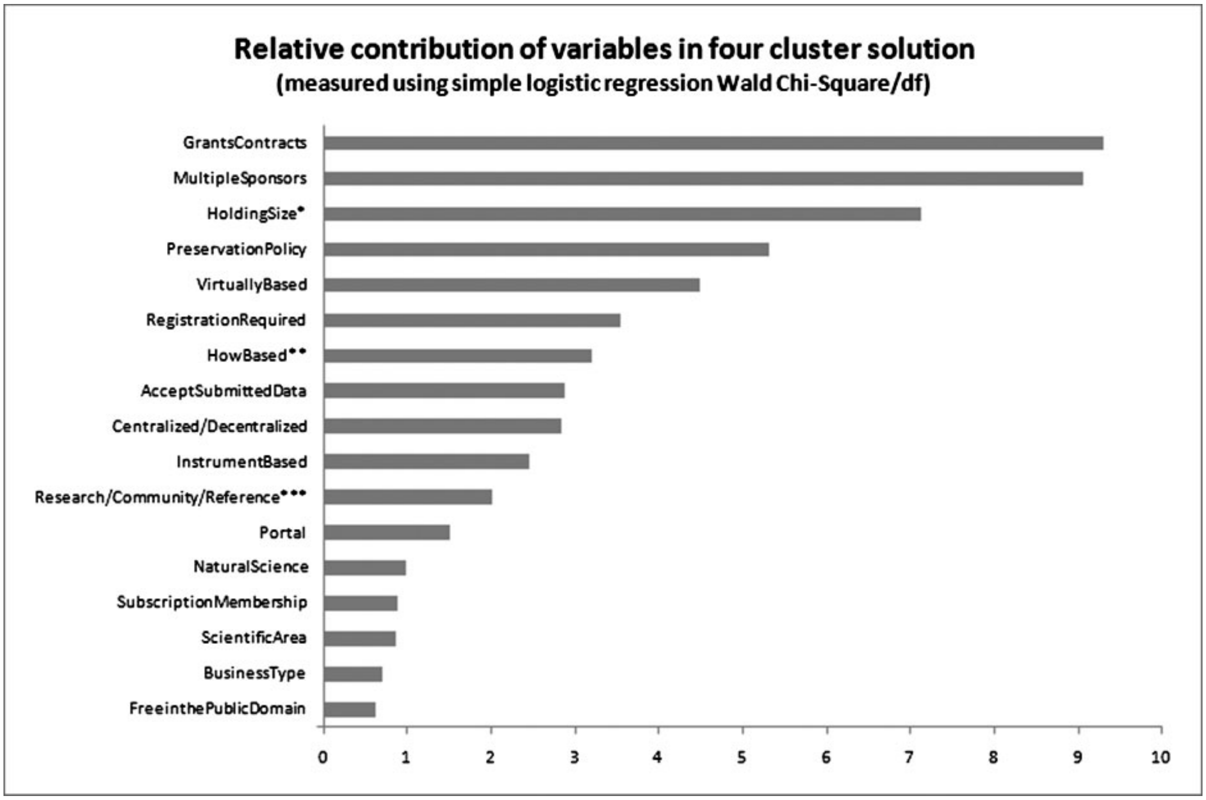


FIG. 4. Relative contribution of 17 analyzed variables to the four-cluster solution as shown in Figure 3. The y-axis value is the Wald chi-square divided by the degrees of freedom for each regressed variable.

*The HoldingSize variable most closely approximates the classification set out in the NSB 2005 report for research, community/resource, and reference level data collections.

**The HowBased variable most closely approximates the distinction set out in the NSB 2005 report between data collections as governmental, university based, or data federations (though here this is broken into the categories independent and aggregate).

***The research/community/reference variable included here, unlike that in the NSB 2005 report, is descriptive of how the overall organization functions. On some level, these are all "research" enterprises, some are particularly "community" centric, and a small number might view themselves as "reference" organizations.

based on expected importance as well as being able to collect reasonably accurate and homogenous data from SDR Web sites, the selection of different variables, or potentially more accurate data, could lead to different results. Several different combinations of variables were tested to evaluate the robustness of this solution (details provided upon request) and review of data collected was performed by at least 61% of SDR administrators contacted. The alternative solutions showed only minor changes in group membership, as would be expected. Furthermore, they did not differ greatly from each other in terms of the semipartial r-squared values, suggesting some stability of the results.

Trends. Without an effective way to longitudinally sample the SDR Web sites over time, and with a small subset having been created before 1985 (22), the remaining 78 SDRs (with an inception date of 1985 or after) were studied in terms of mean variable responses over time. One year (1987) in the period from 1985 to 2008 had no observations. The cutoff date of 1985 was selected, because it marked a time of incredible growth in the development of new technologies like the personal computer, the Internet, and the advent of a variety of informatics disciplines. Although this coarsely samples the time period and may be confounded by nonrandom sampling of the SDRs, it may provide hints as to how some SDR features are changing over time (see details at http://ils.unc.edu/bmh/pubs/SDR_final_sheet.xlsx). Looking closely at the top 10 variables identified as differentiators in Figure 4, SDRs with grant and contract support as well as multiple sponsors is on the increase. This may be because of, in large part, an increased tendency for major governmental agencies like the National Science Foundation (NSF) and the National Institutes of Health (NIH) to provide funding to external entities for projects, perhaps contributing to the trend away from governmentally based SDRs toward more independent and aggregate SDRs. The number of university-based SDRs appears to have remained steady. These findings probably relate to the observation that SDRs in the holding size category of 2 = medium/broad appear to be increasing, along with perhaps an increasing trend in observed 1 = small/less broad SDRs and a steady emergence of 3 = large/more broad type SDRs. There is also some

TABLE 7. The four groups as represented in Table 6 are presented.

	Cluster A:	Cluster B:	Cluster C:	Cluster D:
Variables	Governmental	Medicine/Small	University	Community Biology
GrantsContracts	No	Mixed	Yes	Yes
MultipleSponsors	No	Yes	Yes	Mixed
HoldingSize	Large	Small	mixed	Moderate
PreservationPolicy	Yes	Mixed	Yes	No
VirtuallyBased	No	Mixed	No	No
RegistrationRequired	No	No	Mixed	No
HowBased	Governmental	Mixed	University	Mixed
AcceptSubmittedData	Mixed	Yes	Yes	Yes
Centralized/Distributed	Mixed	Mixed	Mixed	Distributed
InstrumentBased	Mixed	No	No	No
Research/Community/Reference	Research	Mixed	Research	Community
Portal	Mixed	No	Mixed	Mixed
NaturalScience	Yes	Yes	Yes	Yes
SubscriptionMembership	No	No	No	No
ScientificArea	Mixed	Medicine	Mixed	Biology
BusinessType	Federal Center	Mixed	University	Partnership
FreeinthePublicDomain	Yes	Yes	Yes	Yes

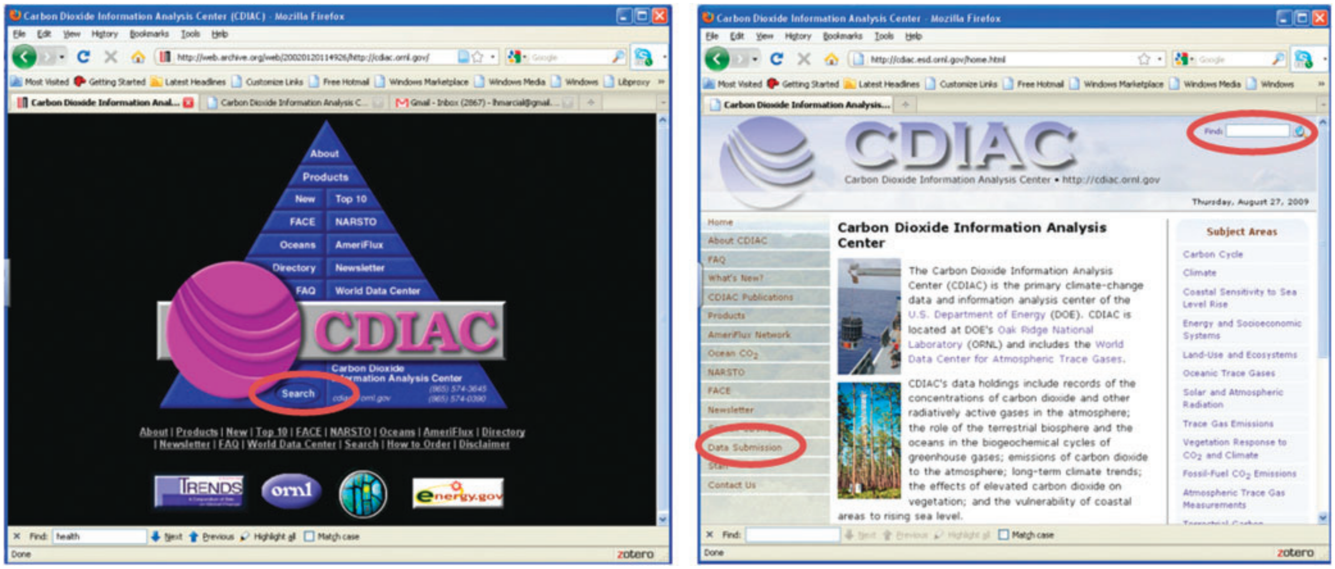
Note. Shading connotes noteworthy differences by variable in comparison to the other groups.

evidence that distributed SDRs, those "housed in a set of physical locations and linked together electronically to create a single, coherent collection" (NSB, 2005), appear to be on the rise.

The existence of a preservation policy, while fairly steady over the period, appeared to fluctuate a little, and it is unclear whether this is represents a real pattern or an artifact of the data. There may be some evidence of a decline in SDRs, which consider themselves primarily virtually based. It could be presumed that this finding is the result of a trend toward the procurement of dedicated staffing or organizational infrastructure to assist with data curation, stewardship, and management, changing the nature of an otherwise virtually based organization. Two other important rising trends are that of a registration requirement, limiting use in some cases to subscribers or members but in more general use to help track usage of the data collections and other services of the SDRs. There is also perhaps a slight increase in the tendency to both provide data for export as well as ingest.

Longitudinal analysis. Based on a temporal review of the Web site for an SDR, there is a lack of clear beginning and end points on the life cycle of an SDR and little information on SDRs that may have existed but, for one reason or another, did not remain in existence. It is critical to be able to observe the Web presence of SDRs over time to be able to track changes in characteristics, as well as potential measures for success or failure. The Internet Archive, which captures and archives Web pages over time, could be used for such a comparison. To investigate the feasibility of this approach, more than a dozen SDRs included in the study were investigated using historical Web site data from the Internet Archive (1997–2007). The basic principle of the archive is to attempt to capture "snapshots" of Web sites by URL over time.

Figure 5 shows a comparison of the Carbon Dioxide Information Analysis Center (CDIAC) Web site, using the Internet Archive Wayback Machine to obtain an archive taken January 20, 2002, and the current "live" snapshot taken September 24, 2009. Note how the older site is less user friendly, less ADA compliant, and more graphically intensive. Importantly, the search has changed over time from being a node to being a central and persistent feature in the top right navigation. Notice also how the navigation has been developed into a sophisticated structure, aimed at repeat use and displaying the wealth of information underlying the site. Of particular interest, the presence of a "Data Submission" element is present now, which didn't appear in the original version. Despite these changes, this Web site demonstrates more consistency than the bulk of the sites investigated, using the Wayback Machine. For a number of the sites examined, data could not be retrieved for a variety of reasons, including data retrieval failures, sites using tools that (perhaps inadvertently) block the archival process, inconsistent URLs over time, incomplete data, and, in many cases, the visual comparisons were unclear, as in the case of the CDIAC illustration, because of unavailable images or comparison tool limitations. The Wayback Machine also does not necessarily capture many layers worth of information from a Web site and problems with fixity for newer sites with more sophisticated back end programming and dynamically driven pages are apparent. The potential to use the Wayback Machine for longitudinal analysis of SDRs is exciting, but, for this study, it was not possible to perform adequate comparisons using this methodology.



Carbon Dioxide Information Analysis Center (CDIAC) - Mozilla Firefox
http://web.archive.org/web/20020120114926/http://cdiac.ornl.gov/
About
Products
New
Top 10
FACE
NARSTO
Oceans
AmeriFlux
Directory
Newsletter
FAQ
World Data Center
CDIAC
Carbon Dioxide Information Analysis Center
Search
About | Products | New | Top 10 | FACE | NARSTO | Oceans | AmeriFlux | Directory | Newsletter | FAQ | World Data Center | Search | How to Order | Disclaimer
TRENDS
ornl
energy.gov
Find: health
Carbon Dioxide Information Analysis Center - Mozilla Firefox
http://cdiac.esd.ornl.gov/home.html
CDIAC
Carbon Dioxide Information Analysis Center • http://cdiac.ornl.gov
Find:
Thursday, August 27, 2009
Home
About CDIAC
FAQ
What's New?
CDIAC Publications
Products
AmeriFlux Network
Ocean CO2
NARSTO
FACE
Newsletter
Data Submission
Contact Us
Subject Areas
Carbon Cycle
Climate
Coastal Sensitivity to Sea Level Rise
Energy and Socioeconomic Systems
Land-Use and Ecosystems
Oceanic Trace Gases
Solar and Atmospheric Radiation
Trace Gas Emissions
Vegetation Response to CO2 and Climate
Fossil-Fuel CO2 Emissions
Atmospheric Trace Gas Measurements
zotero

Carbon Dioxide Information Analysis Center

The Carbon Dioxide Information Analysis Center (CDIAC) is the primary climate-change data and information analysis center of the U.S. Department of Energy (DOE). CDIAC is located at DOE's Oak Ridge National Laboratory (ORNL) and includes the World Data Center for Atmospheric Trace Gases.

CDIAC's data holdings include records of the concentrations of carbon dioxide and other radiatively active gases in the atmosphere; the role of the terrestrial biosphere and the oceans in the biogeochemical cycles of greenhouse gases; emissions of carbon dioxide to the atmosphere; long-term climate trends; the effects of elevated carbon dioxide on vegetation; and the vulnerability of coastal areas to rising sea level.

FIG. 5. Sample use of the Internet Archive WayBack Machine to compare an archive taken January 20, 2002, and a "live" snapshot taken August 27, 2009, for the Carbon Dioxide Information Analysis Center.

Sustainability and Success

In a 2008 report on online academic resources (OARs) from the Ithaka Foundation (Maron, 2008), sustainability is defined as "having a mechanism in place for generating, or gaining access to, the economic resources necessary to keep the intellectual property or the service available on an ongoing basis." To this end, this study attempted to capture a number of different as well as interrelated variables on business type, organizational structure, funding sources, and strategies as well as information about repository holdings and data handling and management practices. A Web site snapshot may seem an unusually cursory way to approach such a complex problem, but one could argue that when a user seeks out organizations with which he or she wishes to exchange data, a Web site is a natural place to begin. If this kind of information is not readily available or ascertained from a public Web site, perhaps many potential users of the SDR will seek alternative sources or other ways to differentiate among them. In addition, this sampling method does not necessarily differentiate SDRs that might have been in existence primarily as a proof of concept with no real intention of maintaining access over time. It also did not capture those SDRs that do not necessarily maintain a Web presence. Although the field of SDRs is growing rapidly and no single area appears to be saturated at this point, changes in related business environments like print publishing suggest that the landscape is or will become increasingly competitive (Eisen, 2009).

The same Ithaka report (Maron, 2008) goes on to say that "In our opinion, delivering impact is the key factor in the potential for achieving long-term sustainability; only high impact and highly useful materials will draw the financial support from beneficiaries needed for long-term success." In this study, success was difficult to measure. For one, there remains no good working definition of a successful SDR and no clear identification of *measurable characteristics* that might help in making comparisons (Schmitz, 2008). Characteristics that were considered included growth of the SDR in terms of data sets held, number of ingests and exports of datasets, size and detail level of Web presence, and existence on the Web. Disappearing from the Web was assumed to equate with failure. These success characteristics could rarely be extracted with any accuracy via Web searches, although possibly via the SDR administrators through a much more in depth analysis. An important metric, the disappearance of Web sites could be seen only in a few cases. As a result, it was difficult to draw any conclusions about the success or failure of SDRs.

The most promising feature for capturing success may be usage statistics. There is an increasing trend towards the capture of usage data, including Web site statistics, size of collections, data requests, and service requests. Additionally, in a few cases like GenBank, DDBJ, and EMBL, use relative to a referential collection "universe" is captured as well. At present, usage statistics appear to be nonstandardized and are mostly an afterthought or used only for internal purposes. These findings are in accord with the Ithaka report (Maron, 2008, p.6) conclusion that "The absence of focused effort on use, impact, and competition among these types of projects has deep implications for their potential long-term success."

Discussion

Framework

There is an almost universal recognition that SDRs are critical to the future of science, and a means for preserving them and for leveraging their richness across disciplines is needed. It is less clear what the essential components of SDRs are that will make this possible. The research presented here describes a framework for observing, cataloguing, characterizing, and comparing a widely heterogeneous set of digital resources, which are still not well understood and for which clear long lasting support remains elusive. What emerges in the framework as essential components in understanding SDR composition include funding (GrantsContracts and MultipleSponsors), size or scope of data collection (HoldingSize, which correlates to the NSB 2005 report definitions of research, community, or reference), and the existence of formal policies regarding long-term storage of data (PreservationPolicy). In addition to these, the structure of the SDR (HowBased), the business type (BusinessType), and the scientific area (ScientificArea) are important characteristics of group membership. These can be hard to catalog but are important in demonstrating that the groups are differentiable along fairly clear lines.

The NSB 2005 report defined an important characteristic bundling holding size/type and funding as follows. *Research:* products of one or more focused research projects and typically contain data that are subject to limited processing or curation. These collections are generally small or project specific. *Community data collections*: serve a single science or engineering community. They are generally intermediate in size and supported in a somewhat more distributed fashion by the community served. *Reference data collections*: serve large segments of the scientific and education community. These are generally broad and/or multidisciplinary as well as long lived.

These working definitions of a clearly important set of characteristics have been used in the literature to help create a typology of SDRs/data collections. This study examined a characteristic labeled HoldingSize, which mapped almost directly to the "collection size" component of the NSB report's classification criteria. In this study, HoldingSize was not associated with funding/business type when collected.

Strong responses from a few SDR administrators, which indicated that capturing a single value for HoldingSize was inappropriate, made it clear that a framework for SDRs centered on size could be limiting. This was true whether the characteristic measured holdings, scientific depth or breadth, size of user community, funding, staffing, or infrastructure. Based on our data and communication with SDR administrators, one variable alone, even one that sensitively bundles multiple characteristics, does not adequately type the landscape. Moreover, without separating some of these characteristics, it is difficult to make comparisons across SDRs of different types. Several SDR administrators indicated that the use of holding size inappropriately or insufficiently classified their resource, demonstrating that ill-fitting characteristics can impose limitations on the perceived value of an SDR. The real value of this search and discovery approach is in not having a preconceived notion of how SDRs are structured or even how they should be. By allowing their inherent characteristics to emerge from the data, perhaps there is better observation of them. Like library collections, domain breadth or depth, collection novelty, user community as a function of the total user community, and funding variety and breadth all play an important role in differentiation. In the case of the LTER, a self-description might be described as more moderate and broad while also being community-centered. This was not readily evident through a review of the Web site, demonstrating the breadth and sustainability of a large, very broad and influential organization that might be categorized at the "reference" level according to the NSB report (2005) definition. To differentiate it more fully, this review of the LTER site indicated that the LTER operates through member organizations, which helps to give a more adequate characterization.

Additional factors may be needed to further describe the complexity of SDR holdings. Examples might include the sheer size of individual files (from petabytes to megabytes), number of files, facets of storage, and retrieval systems like the ability to store and retrieve within a single database or in a data grid. Also important is whether the SDR supports a (or any) combination of archiving, data picking, and data streaming (Rajasekar, Wan, Moore, Kremenek, & Guptil, 2003). Although the SDRs observed in this study were hard to classify in these terms from a Web site review, it can be presumed that the number of SDRs will continue to grow. As the tendency to capture data from sensors in real time increases and the availability of sophisticated storage and retrieval systems increases, so will SDR growth. From a practical standpoint, this means that future SDRs will be more likely to support data streaming direct from the data creation environment. This will require the provision of access to these streams through pipelines or data grid registration systems. It will also necessitate the development of sophisticated mechanisms for management and access of these data and their multifaceted provenance.

As pointed out by Palmer et al. (2007), there is undoubtedly a "long tail" element to the SDR framework. What is clear is that development of a framework affords a better understanding of the wide variety of SDRs "in the wild" and that this, in turn, should improve measures of sustainability or success. It will be critical, though, to develop meaningful, standard metrics for evaluating SDR success and be able to track SDRs longitudinally. The NSB report (2005) points out: "The distinction between centralized and distributed collections can have important implications for developing policy for funding and for ensuring their persistence and longevity."

The Broader Environment for SDRs

The SDRs reviewed here appear to have started with an idea that resulted from a need for information/data exchange, storage, or management. The ideas that actually culminated in an SDR appear to have had strong vision and leadership, at least strong enough to mature to the next stage. As they mature, issues of funding sources, scope, and possible contributors are decided. Though these decisions may be revisited at any point in this evolutionary process, they form the basis for the public view of the SDR at this emergent stage. The next step appears to be an investment in technology and services to support the concept. This is followed by a service offering stage, perhaps the most widely variable

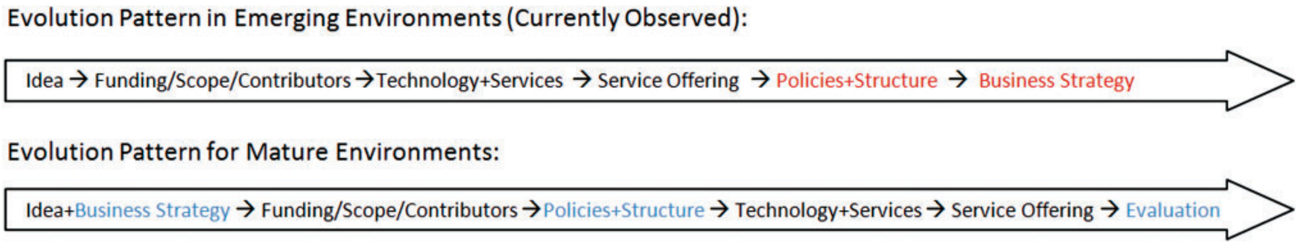

Evolution Pattern in Emerging Environments (Currently Observed):
Idea → Funding/Scope/Contributors → Technology+Services → Service Offering → Policies+Structure → Business Strategy
Evolution Pattern for Mature Environments:
Idea+Business Strategy → Funding/Scope/Contributors → Policies+Structure → Technology+Services → Service Offering → Evaluation

FIG. 6. Current environment and mature environment evolution pattern for SDRs.

of all stages, which is highly dependent upon scope. What appear to be developing next are the policies that guide data exchange among contributors and an organizational structure that can support the growing user base. An area of increasing interest that appears to be emerging is that of a formalized business strategy. This is particularly important for both existing and future SDRs, because although the business case for these entities has yet to be clearly made, their significance is intuitively clear (President's Information Technology Advisory Committee [PITAC], 2005; NSB, 2005; IWGDD, 2009) and information about their availability and use has become increasingly requested and even required.

As is discussed in great detail in the Ithaka report (Maron, 2008), many OARs currently emerge very differently from typical business entities, which are characterized by sales and marketing cycles and early business-oriented strategic development and planning. This may be, in part, because of the fact that they often emerge from governmental, university, or other nonprofit entities that are charged with and oriented toward making their data publicly available. Until recently, these entities have functioned devoid of the pressures most businesses feel to remain solvent. That is definitely changing and a call is out (IWGDD 2009) to investigate how SDRs have emerged and to draw comparisons to business organizational theory that might offer insight into how related organizations have emerged. This might help identify key elements that may ensure the future success of SDRs. The general pattern of evolution observed in this study on SDRs is noted in Figure 6 along with a proposed 'Future' pattern.

As is shown in Figure 6, it is presumed that the evolutionary model will change in the coming years as government entities realize the need to find sustainable funding to maintain these repositories and as the focus on economies of scale in this arena increases. As a result, as more SDRs emerge, traditional market forces will come into play, resulting in more formalized business structures. This will make the identification of business strategy as important as the initial idea and will force the development of policies and structure at an earlier stage, a key differentiator in a saturated market. Difficult to analyze at the "species identification stage," the development of an evaluation and refinement phase will help inform growth and development in these latter stages of SDR evolution.

Another critical element of the SDR life cycle is the general recognition that rather than just making information available via the Internet, a sophisticated ecological framework is emerging in support of these repositories. In reference to the structure of the Global Biodiversity Information Facility (GBIF), the IWGDD report (2009) notes: "This breadth of participation and collaboration provides a potential foundation for sustainability analogous to that provided by diversity in ecosystems sustainability." This ecological analogy more aptly describes the "living" nature of SDRs and of the continuous life cycle of the data they shepherd.

SDRs are conceptually similar to IRs. In Table 8, comparisons between SDRs and IRs are described based on the characteristics identified in this work. The comparison is complicated by the nature and maturity of SDRs, which overwhelmingly precede the emergence of institutional repositories. In addition, SDRs have long-standing communities of practice and have been employing sophisticated infrastructures for years to support their operations. Although the very different nature of SDRs and IRs in terms of domain specificity, and to some extent utility, make the analogy even more complex, there are also overwhelming similarities in structure, operations, and use as seen in their common characteristics shown in Table 8.

Recommendations for Development and Success of SDRs

Several recommendations to the current and potential organizers of SDRs have developed in the course of this study. This study attempted to obtain information that would be of interest to any SDR user or contributor. In the course of data collection, information was often either difficult to find or simply not available on the Web. Another issue was standardization. Though an important strength of an SDR may be its domain specificity, it is increasingly apparent that modern scientists find themselves working across domains to solve problems. This interdisciplinary work is greatly facilitated when an adopted standard creates an environment for open and easy data sharing. This has happened with institutional repositories and the development of several commonly used freely available software platforms (DSpace, Eprints, Fedora, and Greenstone). It is possible that once the common elements of SDRs are recognized, a freely available open source platform that supports SDRs might be developed. This could help lower the barrier to entry for emerging SDRs, as well as encourage good management practice. The cross-sectional approach taken here may result in a better overall awareness of existing practices among

TABLE 8. A comparison of institutional repository development and SDR development—a review of the literature.

Characteristic	Institutional Repository	Science Data Repository
Holdings management	IRs have a high degree of similarity in terms of management of holdings.	SDRs are dissimilar, often highly domain specific, to each other in terms of holdings.
Handling Procedures	Homogeneity of handling procedures both within and among repositories (DRIVER, 2008)	Heterogeneity of handling procedures, perhaps necessary to degree of specialization within a domain, often seemingly due to lack of standardization
Base	Institutionally based (DRIVER, 2008)	Typically domain based, though increasingly cross-cutting, making the call for standardization more critical
Evolutionary stage	Middle stage of evolution (Robertson et al., 2007)	Early, "species identification" stage of evolution as a wide variety of researchers contribute to the development of a "typology" (NSB, 2005; IWGDD, 2009)
View	Macro view: IRs typically function at a high level incorporating media of a wide variety of types across domains with the overarching goals of long term storage and interoperability.	Typically these originate with a micro view within a domain. For the largest SDRs, either the domain breadth expands; specialization increases or the SDR takes a more interdisciplinary approach to support its user community.
Degree of specialization	Attempting to aggregate the highly specialized (Lynch, 2003)	Highly specialized (Lynch, 2008)
Cosmic view model For more details see: http://www.rubric.edu.au/extrafiles/wheel/main.swf	Cosmic view model: For IRs this model is effective in flexibly describing the relevant layers of characterizing a given IR (Blinco & McLean, 2004)	Cosmic view model: For SDRs this model lacks the critical domain layer that these entities are often defined by. It also insufficiently covers the problem space of sustainability for SDRs tied directly to business characteristics.
Grounding	Assumed to be grounded in an institution (DRIVER, 2008)	Cannot assume grounding in a single institution
Business structure	Business structure still evolving but outwardly appears somewhat similar (DRIVER, 2008)	High degree of variability in business structure, directly affecting issues of sustainability and success.
Characteristics of success/group composition	Important characteristics of successful repositories (DRIVER, 2008): • Business of digital repositories • Stimuli for depositing materials into repositories intellectual property rights • Data curation • Long-term preservation	Pivotal characteristics of SDR groups: • GrantsContracts • MultipleSponsors • HoldingSize • PreservationPolicy

SDRs, regardless of size, breadth, or impact. This agrees with the assertion made by Borgman et al. (2007) regarding "little science" benefiting from "big science" in the development of both guidelines and tools. For example, an important advantage for smaller, grassroots SDRs is that they can be aggregated through larger portal-type sites like geodata.gov (http://gos2.geodata.gov/wps/portal/gos). Making it easier to discover individual SDRs as well as information about them may also encourage adherence to common metadata standards. An additional benefit may be an exchange in the reverse direction where big science SDRs learn from experimentation and innovation occurring in little science SDRs. Following these data longitudinally would improve our ability to define and measure SDR success.

Efforts are underway to encourage repository managers to improve information sharing and access. Increasing metadata creation and standardization, use of open source tools or of more extensible tools, and the development of export, ingest, curation, archival, and preservation and storage policies are just some of the methods observed among a subset of SDRs. It is unclear whether domain specificity and data handling have precluded many SDRs from broader adoption of these policies but perhaps this plays a role. Although this study has made an effort to disambiguate the SDR environment, it also renews an acknowledgement of the fact that heterogeneity may continue to be a challenge in defining a framework for SDRs.

Technology has made it easier to develop or start a SDR, but as is evident from this study, a lot of effort is still required to maintain them. As a result, SDRs without substantial investment in infrastructure and support do not survive and thrive. It is not clear whether governmental funding is critical to SDR sustainability, but it would appear that it might be critical to success at the emergent stage and that a large subset of the SDR landscape remains directly government supported. Although the federal government has recognized that it should support scientific data repositories (PITAC, 2005; NSB, 200; IWGDD, 2009), the primary support has been limited to NSF funding initial development of repositories. And although there are requirements for sharing of data and data management preparation in advance of receiving funding (NIH, 2010; IWGDD, 2009), it is less clear where long-term, sustainable funding for SDRs will come from. It can be anticipated, that SDRs should increasingly develop their own funding models to be successful (Maron, 2008; Maron et al., 2009). Exceptions to this might continue to be large governmentally funded repositories like Genbank, where recognition of the need to promote data sharing plays a pivotal role in sustainability. A recent positive step is the NSF's Sustainable Digital Data Preservation and Access Network Partners (DataNet) Program, which is providing substantial funding to five partners to create "exemplar national and global data research infrastructure organizations," with the explicit goal to build sustainable infrastructure.

SDRs play a critical role in the future success of science. This study provides a baseline survey of their current state, and it highlights an inferred framework for studying SDRs and evaluating their success. These results can help us understand the SDR environment better and provide guidance to SDRs and funders of SDRs, with the hope of making SDRs an established and fruitful part of our global scientific efforts.

References

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., & Cherry, J.M., et al. (2000). Gene ontology: Tool for the unification of biology. Nature Genetics, 25(1), 25.

Blinco, K., & McLean, N. (2004). A 'cosmic' view of the repositories space (wheel of fortune). Retrieved July 16, 2010, from http://www.rubric.edu.au/extrafiles/wheel/main.swf

Borgman, C., Wallis, J., & Enyedy, N. (2007). Little science confronts the data deluge: Habitat ecology, embedded sensor networks, and digital libraries. International Journal on Digital Libraries, 7(1), 17–30.

Carbon Dioxide Information Analysis Center. Internet Archive Wayback Machine: A search on http://cdiac.ornl.gov. Retrieved July 16, 2010, from http://Web.archive.org/Web/*/http://cdiac.ornl.gov/

Chan, L., & Zeng, M. (2006). Metadata interoperability and standardization: A study of methodology. Part I: Achieving interoperability at the schema level. D-Lib Magazine, 12(6). Retrieved July 16, 2010, from http://www.dlib.org/dlib/june06/chan/06chan.html

Cochrane G.R., & Galperin, M.Y. (2010). The 2010 nucleic acids research database issue and online database collection: A community of data resources. Nucleic Acids Research, 38(D1–D4).

David, P.A. (2004). Towards a cyberinfrastructure for enhanced scientific collaboration: Providing its "soft" foundations may be the hardest part. Retrieved July 16, 2010, from http://129.3.20.41/eps/le/papers/0502/0502002.pdf

Digital Repository Infrastructure Vision for European Research. (2008). DRIVER: Networking European scientific repositories. Retrieved July 16, 2010, from http://dare.uva.nl/document/150724

Eisen, B. (2009). Digital – and financially viable. Inside Higher Ed. Retrieved July 16, 2010, from http://www.insidehighered.com/layout/set/print/news/2009/07/15/ithaka

Garritano, J.R., & Carlson, J. (2009). A subject librarian's guide to collaborating on e-science projects. Retrieved July 16, 2010, from http://www.istl.org/09-spring/refereed2.html

Interagency Working Group on Digital Data. (2009). Harnessing the power of digital data for science and society. Report to the Committee on Science of the National Science and Technology Council. Retrieved July 16, 2010, from http://www.nitrd.gov/about/Harnessing_Power_Web.pdf

Johnson, S.C. (1967). Hierarchical clustering schemes. Psychometrika, 32(3), 241–254.

Karasti, H., & Baker, K.S. (2008). Digital data practices and the long term ecological research program growing global. International Journal of Digital Curation, 3(2), 42–58.

Lynch, C. (2003). Institutional repositories: Essential infrastructure for scholarship in the digital age. Association of Research Libraries, 226, 1–7.

Lynch, C. (2008). Big data: How do your data grow? Nature, 455(7209), 28–29.

Madrigal, A. (2008). Google to host terabytes of open-source science data. Retrieved July 16, 2010, from http://www.wired.com/wiredscience/2008/01/google-to-provi/

Marchionini, G. (1998). Consider a sharium. Retrieved July 16, 2010, from http://www.ils.unc.edu/∼march/sharium/sharium1.1.html

Maron, N. (2008). Sustainability and revenue models for online academic resources. An Ithaka Report. Retrieved July 16, 2010, from http://www.ithaka.org/strategic-services/sca_ithaka_sustainability_report-final.pdf

Maron, N.L., Smith, K.K., & Loy, M. (2009). Sustaining digital resources: An on-the-ground view of projects today. An Ithaka Report. Retrieved July 16, 2010, from http://www.ithaka.org/ithaka-s-r/strategy/ithaka-case-studies-in-sustainability/report/SCA_ithaka_SustainingDigitalResources_Report.pdf

National Coordination Office for Information Technology Research and Development. (May 27, 2005). PITAC 2005: President's Information Technology Advisory Committee June 2005 report to the president on "Computational science: Ensuring America's competitiveness." Retrieved July 16, 2010, from http://www.nitrd.gov/pitac/reports/20050609_computational/computational.pdf

National Science Board [NSB]. (2005): Long lived digital data collections: Enabling research and education in the 21st century, National Science Board (NSB-05-40, Revised May 23, 2005). Retrieved July 16, 2010, from http://www.nsf.gov/pubs/2005/nsb0540/

National Institutes of Health Office of Extramural Research. (2007). NIH data sharing policy. Retrieved July 16, 2010, from http://grants.nih.gov.libproxy.lib.unc.edu/grants/policy/data_sharing/

Nature. (2008, September). Big data. Nature, 455, 7209.

Palmer, C.L., Cragin, M.H., Heidorn, P.B., & Smith, L.C. (2007, December). Data curation for the long tail of science: The case of environmental sciences. Paper presented at the Third International Digital Curation Conference, Washington, DC.

Rajasekar, A., Moore, R., Wan, M., & Schroeder, W. (2009). Universal view and open policy: Characteristics of large-scale data management system. Proceedings of the 2009 International Symposium on Collaborative Technologies and Systems (pp. 322–329). Washington, DC: IEEE Computer Society.

Rajasekar, A., Wan, M., Moore, R., Kremenek, G., & Guptil, T. (2003). Data grids, collections, and grid bricks. Proceedings of the 20th IEEE Symposium on Mass Storage Systems and Eleventh Goddard Conference on Mass Storage Systems and Technologies. Washington, DC: IEEE Computer Society.

Robertson, R.J., Mahey, M., & Allinson, J. (2007). An ecological approach to repository and service interactions. Retrieved July 16, 2010, from http://www.ukoln.ac.uk/repositories/digirep/images/a/a5/Introductoryecology.pdf

Romesburg, H.C. (2004). Cluster analysis for researchers. Belmont, CA: Lifetime Learning Publications.

Schmitz, D. (2008). The seamless cyberinfrastructure: The challenges of studying users of mass digitization and institutional repositories. Retrieved July 16, 2010, from http://www.clir.org/pubs/archives/schmitz.pdf.

Wulf, W.A. (1989, March). The national collaboratory—A white paper (Appendix A). Unpublished report of a National Science Foundation invitational workshop held at Rockefeller University, New York.