DATA SCIENCE JOURNAL

Assante, M et al 2016 Are Scientific Data Repositories Coping with Research Data Publishing? *Data Science Journal,* 15: 6, pp. 1–24, DOI: http://dx.doi.org/10.5334/dsj-2016-006

REVIEW

Are Scientific Data Repositories Coping with Research Data Publishing?

Massimiliano Assante[1], Leonardo Candela[1], Donatella Castelli[1] and Alice Tani[1]

[1] Istituto di Scienza e Tecnologie dell'Informazione "Alessandro Faedo", Consiglio Nazionale delle Ricerche, Via G. Moruzzi 1, 56124, Pisa, Italy
leonardo.candela@isti.cnr.it

Corresponding author: Leonardo Candela

Research data publishing is intended as the release of research data to make it possible for practitioners to (re)use them according to "open science" dynamics. There are three main actors called to deal with research data publishing practices: researchers, publishers, and data repositories. This study analyses the solutions offered by generalist scientific data repositories, i.e., repositories supporting the deposition of any type of research data. These repositories cannot make any assumption on the application domain. They are actually called to face with the almost open ended typologies of data used in science. The current practices promoted by such repositories are analysed with respect to eight key aspects of data publishing, i.e., dataset formatting, documentation, licensing, publication costs, validation, availability, discovery and access, and citation. From this analysis it emerges that these repositories implement well consolidated practices and pragmatic solutions for literature repositories. These practices and solutions can not totally meet the needs of management and use of datasets resources, especially in a context where rapid technological changes continuously open new exploitation prospects.

Keywords: Research Data Publishing; Scientific Data Repositories; Data infrastructures; Data Quality

1 Introduction

Research data publishing (Klump et al., 2006; Lawrence et al., 2011; Costello et al., 2013) is an approach for sharing research data, i.e., it is intended as the release of (research) data for (re)use by others. This approach is usually based on the publishing metaphor (Parsons and Fox, 2013; Borgman, 2015) and this contributes to controversies and debates (Kratz and Strasser, 2014). In this work the term "data publishing" is meant as the act of making a *dataset* – data treated collectively as a unit (Renear et al., 2010) – public thus to enable its dissemination and (re)use, i.e., it includes but does not inherit the full meaning it generally is given in the area of scientific and technical publishing. The motivations underlying the publishing act lie in the willingness to overcome the methodological, legal, and technical barriers preventing research data sharing to be the norm in science (Borgman, 2011; Douglass et al., 2014; Asher et al., 2013; Bourne et al., 2012; Bourne, 2010; Tenopir et al., 2011; Pampel and Dallmeier-Tiessen, 2014).

Scientific data repositories (Marcial and Hemminger, 2010) have a key role in science. They are called to implement systematic data stewardship practices thus to foster adequate scientific datasets collection, curation, preservation, long term availability, dissemination and access. Such repositories are largely diffused whitin communities that produce a huge amount of data, such as physics (e.g., SDSS at FermiLab), genetics (e.g., GenBank at NCBI Data), or environmental sciences (e.g., British Atmospheric Data Centre). Recently, the demand for data repositories is emerging also within the so-called "*long-tail*" of science, i.e., in the context of those scientific domains in which activity is performed in a large number of relatively small labs and by individual researchers who collectively produce the majority of scientific results (Heidorn, 2008). In such contexts the research conducted is specific, diverse, cutting-edge and traditionally lacks shared community repositories (Borgman, 2015). The actual needs of sciences that rely on smaller "research level" data

collections are at the moment less understood and cannot completely be based on experiences and models made within standardised disciplinary repositories (Palmer et al., 2007).

Scientific data repositories are often proposed as instruments for supporting data publishing as they provide facilities for all the different players involved in this process.

Despite the pressing need for proper research data publishing practices and the proliferation of scientific data repositories as means contributing to these practices there is no shared understanding of what repositories should indeed offer to support data publishing. By analyzing the current solutions offered by existing repositories, this survey presents a systematic review of the state of the art with the aim of identifying gaps and suggest directions for improvements.

This survey focuses on "generalist" scientific data repositories, i.e., repositories that make no assumptions nor special arrangements for community- or data-type-specific aspects. They are thus open to publish any type of research dataset. These kinds of repositories are especially developed to support the publishing of datasets produced in the "long-tail" scientific contexts because of the heterogeneity of the possible outcomes.

This study complements a previous one on the approaches for data publishing promoted by publishers, i.e., data journals (Candela et al., 2015).

The paper is organised as follows. Section 2 discusses the criteria used in selecting the repository sample to be analyzed and briefly introduces each of the identified "generalist" repository. Section 3 gives an overview of the datasets published in the selected repositories up to 2015. Section 4 analyses the selected repositories with respect to the eight aspects they are called to cope with for data publishing, i.e., dataset formatting, documenting, licensing, publication costs, validation, availability, discovery and access, and citation. For each of these aspects the study discusses the current practices and their limitations. Finally, Section 5 concludes the article by giving suggestions aiming at reinforcing the entire research data publishing ecosystem.

2 Repository Selection

Directories of research data repositories[1] enumerate a very large and continuously increasing number of repositories, most of them being disciplinary ones with very specific characteristics. Currently, only a handful of them are "generalist" ones. However, the demand for generalist repositories is quickly growing because of the many forces (e.g., governments, funding agencies, journals) pushing for systematic management and sharing of research data. Thus, a generalist data repository is required whenever a scientific community has no reference data repository for deposition of the type of datasets is is producing.

In this survey we will focus on generalist scientific data repositories recommended by data journals for the deposition and publication of research data (Candela et al., 2015). This choice enables us to affirm that the selected repositories have a role in the data publishing practice already acknowledged by the research community.

This selection process has led to the identification of the following five repositories[2] that constitute the core pool analysed in detail by this study.

3TU.Datacentrum. The repository[3] originated from the cooperation of the three technical universities in the Netherlands (Rombouts and Princic, 2010). Its aim is to provide the scientific community with a sustainable and persistent archive for research data. Its management is supported by the TU Delft Library.

CSIRO Data Access Portal. The repository[4] was established by the Australian national science agency CSIRO (Commonwealth Scientific and Industrial Research Organization). It aims to manage, discover and share data across different research fields. It is part of the data services under the umbrella of the Australian National Data Service (ANDS).

Dryad. The repository[5] originated from the initiative of a group of journals aiming to adopt a joint data archiving policy (JDAP) for their publications, as well as to set up a community-governed infrastructure for data archiving and management. It is supported by a nonprofit membership organization, including journals and publishers, scientific societies, research institutions and libraries, and research funding organizations.

[1] For instance, Registry of Research Data Repositories http://www.re3data.org/ or Databib http://databib.org/.

[2] Dataverse (Crosas, 2011), often cited as a generic data repository, is not part of this list simply because it is a technology enabling to create repositories.

[3] 3TU.Datacentrum Website http://datacentrum.3tu.nl

[4] CSIRO Data Access Portal https://data.csiro.au

[5] Dryad Website http://datadryad.org/

Figshare. The repository[6] was started by Mark Hahnel, an Imperial College PhD student passionate about open data, as a way to store, manage and freely disseminate any kind of research output. It is funded by Digital Science, the global technology division of Macmillan Science & Education. This division invested in a series of initiatives aiming at providing software and technology tools to scientists and researchers, e.g., the Altmetric service (Adie and Roe, 2013), one of the services dealing with measuring the impact of research products by non-traditional metrics (Piwowar, 2013).

Zenodo. The repository[7] was launched within the EU FP7 project OpenAIREplus (Manghi et al., 2012) as part of a European-wide research infrastructure. It aims to enable researchers to preserve and share any kind of research output, with a particular focus on those produced in the context of the long-tail of science. The repository is co-funded by the European Commission via OpenAIRE and hosted by CERN. It recently launched a crowdfunding campaign for expanding features and storage capabilities, e.g., with a donation of CHF 2500 it is supported 1 week of overall service management.

The survey has been performed by analysing the repository websites, by searching the web for literature about them, and by contacting repository responsible(s) when needed. Moreover, repositories' datasets have been systematically collected (actually their metadata) and analysed to derive a characterisation of repositories' content.

Table 1 gives some basic data on the selected repositories including their type, year of foundation, base location, underlying software, and whether the repository is certified or not[8]. Being generalist, these repositories do not have a specific designated community (CCSDS, 2012). In fact, the communities that use each repository are quite diverse each other, e.g., Figshare and Zenodo datasets are less discipline focused than CSIRO ones (cf. **Tab. 4**). This heterogeneity is not expected to impact on the criteria we based our investigation because of the discipline-agnostic nature of the proposed criteria.

Before analysing in detail the support that these repositories offer to research data publishing, we provide an overview of the content of these repositories up to December 2015.

3 Published Datasets: an Overview

Up to now there is neither a shared definition of "research data" nor a shared association between this term and the "dataset" term. Often with "research data" it is intended the very broad and heterogeneous range of materials produced during a research activity. This study embraces the definition of (research) data given by Borgman (2015), i.e., "entities used as evidence of phenomena for the purpose of research or scholarship", and uses "dataset" to refer to the unit of data subject of the data publishing activity, no matter how many files it materialises (Renear et al., 2010). This "dataset" definition includes the term "data package" as adopted by Dryad to mean a set of data files associated with a publication, as well as "dataset" and "fileset" as used by Figshare to indicate data (the former) and a group of multiple files citable as a single object (the latter).

The selected repositories declare to accept almost any dataset underlying a research activity. Mainly, they only impose constraints on the dataset "manifestation", i.e., repositories expect datasets to be provided by means of files, and do not consider data streams nor protocols for acquiring the dataset content at deposition time.

	Type	Founded	Country	Software	Cert.
3TU.Datacentrum	Institution	2008	NLD	In-house	✓[a]
CSIRO DAP	Institution	2011	AUS	In-house	
Dryad	Organization	2008	USA	DSpace	
Figshare	Company	2011	GBR	In-house	
Zenodo	Organization	2013	CHE	Invenio	

Table 1: Scientific Data Repositories studied.
[a]Data Seal of Approval

[6] Figshare Website http://figshare.com
[7] Zenodo Website https://zenodo.org/
[8] Certification aims at guaranteeing that a Repository operates according to established standards, that the repository is "trustworthy". No specific certification mechanism has been developed to assess the trustworthiness of repositories with respect to the needs of research data publishing.

Concerning the content: (*a*) Zenodo and Figshare actually deal with any research product, i.e., these repositories make it possible to publish any research outcome including papers, posters and presentations; (*b*) Dryad only accepts datasets associated with a published article; (*c*) CSIRO DAP and 3TU.Datacentrum offer specific support for certain kind of datasets. In particular, CSIRO DAP supports three specific collections: ATNF Astronomy Observations (pulsar observations from the Australia Telescope National Facility in Parkes), AAHL Microscopy Observations (microscopy images from the Australian Animal Health Laboratory), and Sensor Data (empirical data about the natural world coming from the CSIRO Sensor Networks); while 3TU. Datacentrum manages the datasets belonging to the UNESCO-IHE Institute for Water Education.

An overview of the datasets published by the selected repositories up to December 2015 is given in **Tables 2–4** expressed in terms of number of published datasets, number of published files, and most frequent subjects of the published datasets.

The selected repositories have published a total of 336,647 datasets (**Tab. 2**). The large majority of such datasets – 85% circa – has been published in the last three years, namely 33% circa of datasets have been published in 2013, 29% circa in 2014, and more than 22% in 2015. The large majority of datasets comes from Figshare, this repository holds more than 95% of the total of the items published.

The published datasets contribute a total of 916,874 files to the repositories (**Tab. 3**). In average, each dataset contributes approximately 3 files. However, the distribution is not homogeneous. CSIRO DAP is the repository containing the larger number of files. This is due to the presence of items containing a very large number of files, e.g., "Parkes observations for project P309 semester 1999MAYT" consists of 25,804 files. In CSIRO DAP 1% circa of the datasets contains more that 10,000 files each, 3% circa of the datasets contains more that 5,000 files each, while 6.5% of the datasets contains more than 1,000 files each.

Subjects of the published datasets range from biological sciences to physics and genetics (**Tab. 4**). The repository exposing the largest variety of subjects is Dryad where a total of 19,829 distinct subjects have been used to characterise its 9,676 datasets. However, there is a large use of unique subjects, i.e., subjects specifically used for one dataset only. The number of unique subjects per repository is: 585 in 3TU. Datacentrum (80% circa of its distinct subjects), 1318 in CSIRO DAP (68% circa), 15,158 in Dryad (76% circa), 96 in Figshare (32% circa), and 1550 in Zenodo (78% circa). More than 65% of 3TU.Datacentrum datasets actually lack subjects at all. This is somewhat surprising as it happens in a context where tagging research products with keywords, at the very least, might be expected.

Overall these figures demonstrate, as expected, that the information space of the generalist repositories is largely heterogeneous and variegated. These repositories are called to manage datasets produced in

	up to 2010	2011	2012	2013	2014	2015	Total
3TU.Datacentrum	1692	446	379	345	371	296	*3529*
CSIRO DAP	0	46	62	438	454	418	*1418*
Dryad	493	773	1309	1990	2687	2424	*9676*
Figshare	0	16,929	28,224	108,221	94,223	72,818	*320,415*
Zenodo	99	24	68	43	268	1107	*1609*
Total	*2284*	*18,218*	*30,042*	*111,037*	*98,003*	*77,063*	*336,647*

Table 2: Datasets published by Scientific Data Repositories.

			Per dataset	
	Total	Average	Minimum	Maximum
3TU.Datacentrum	3,833	1	1	18
CSIRO DAP	549,514	388	0	25,804
Dryad	30,814	3	1	96
Figshare	320,922	1	0	443
Zenodo	11,791	7	1	3,963
Total	*916,874*			

Table 3: Files published by Scientific Data Repositories.

Subj.	3TU.Datacentrum	CSIRO	Dryad	Figshare	Zenodo
#1	*n/a* (3141 – 67.2%)	020199 (690 – 6.6%)	*n/a* (606 – 1.2%)	biological sciences (155,196 – 23%)	*n/a* (445 – 10.1%)
#2	dts (37 – 0.8%)	pulsars (423 – 4.1%)	adaptation (529 – 1.1%)	medicine (60,866 – 9.3%)	matriz de datos (234 – 5.3%)
#3	stream water t. (37 – 0.8%)	neutron stars (416 – 4%)	pop. gen. – emp. (429 – 0.9%)	genetics (33,080 – 5.1%)	prog. esta. (231 – 5.2%)
#4	530 physics (23 – 0.5%)	comp. bin. . . . (247 – 2.4%)	speciation (326 – 0.7%)	biotechnology (28,470 – 4.4%)	PowerTAC (138 – 3.1%)
#5	000 comp. sci. (21 – 0.4%)	Australia (229 – 2.2%)	ecological gen. (281 – 0.6%)	ecology (23,398 – 3.6%)	3D (90 – 2%)
#6	apsp (19 – 0.4%)	interstellar . . . (186 – 1.8%)	phylogeography (263 – 0.5%)	biochemistry (21,638 – 3.3%)	archaeology (89 – 2%)
#7	shortest path (19 – 0.4%)	adaptation (126 – 1.2%)	hybridization (222 – 0.5%)	infectious diseases (21,561 – 3.3%)	3D doc. (88 – 2%)
#8	stp (19 – 0.4%)	climate change (117 –1.1 %)	insects (219 – 0.4%)	science policy (21,324 – 3.3%)	caddo (88 – 2%)
#9	stn (19 – 0.4%)	pulsar (112 – 1%)	cons. genet. (217 – 0.5%)	uncategorized (21,079 – 3.2%)	Amer. South. (88 – 2%)
#10	hydra. eng. (19 – 0.4%)	alien plant (108 – 1%)	fish (167 – 0.4%)	cell biol. (18,780 – 2.9%)	web craw. (38 – 0.9%)
Distinct	740	1940	19,829	299	1977

Table 4: Top 10 subjects associated with published dataset. The following subjects are abridged to reduce the table size: 'stream water t.' is 'stream water temperature'; '000 comp. sci.' is '000 Computer science, knowledge & systems'; 'apsp' is 'APSP, all-pairs shortest paths'; 'stp' is 'STP, simple temporal problem'; 'stn' is 'STN, simple temporal network'; 'hydra. eng.' is 'hydraulic engineering'; 'comp. bin. . . . ' is 'compact binaries and/or black-holes'; 'interstellar . . . ' is 'interstellar medium in and around the Milky Way'; 'pop. gen. – empirical' is 'population genetics – empirical'; 'ecological gen.' is 'Ecological Genetics'; 'cons. gen.' is 'Conservation Genetics'; 'cell biol.' is 'Cell Biology'; 'prog. esta.' is 'Programas estadìsticos'; '3D doc.' is '3D Documentation'; 'Amer. South.' is 'American Southeast'; 'web craw.' is 'web crawling'. Some of the CSIRO subjects are classification codes coming from the Australian and New Zealand Standard Research Classification (ANZSRC), e.g., 020199 Astronomical and Space Sciences not elsewhere classified.'dts' stands for 'Distributed Temperature Sensing'.

different contexts, by different teams. This heterogeneity calls for more robust and flexible methods and approaches for serving the data publishing expectations than those offered by community specific repositories. In the next section, it is given account of the adopted approaches, highlighting potential limitations and proposing possible enhancements.

4 Analysis

Our analysis of generalist data repositories has been conducted by examining how repositories deal with eight features that play a key role in characterising the support that these repositories offer to the publishing of research data as envisaged in Section 1. These features are:

- *formatting*, i.e., the arrangement of a dataset according to a certain format to ensure long-term usability of the dataset;
- *documenting*, i.e., complementing a dataset with information to facilitate its retrieval, understanding and (re-use) by both humans and tools ;
- *licensing*, i.e., the characterization of the term of use of the data by third parties' consumers, to enable appropriate and informed dataset (re)uses;
- *publication costs*, i.e., the effort and/or amount of funding to be spent by the owner to publish a data set, to guarantee a fair and open access to it;

- *validation*, i.e., the process applied to assess the "cogency" or "soundness" of the published data, to ensure that available datasets are provided with quality-oriented attributes to assess their fitness for purpose;
- *availability*, i.e., the degree to which the published datasets are at consumer's disposal, to guarantee that practitioners get access to published datasets over time;
- *discoverability and access*, i.e., the facilities enabling practitioners to identify datasets of potential interest and to obtain them;
- *citation*, i.e., the provision of a reference to dataset(s) describing data properties enabling credits and attribution, discover, interlinking, and access. support attribution, reward and repeatability.

These features result from the two main phases characterising research data publishing, i.e., the phase of preparation/submission (datasets have to be formatted, documented, licensed, and the cost of publication has to be covered) and the phase of consumption (datasets have to be assessable[9], available, discoverable and citable). For each of these features the paper discusses: (*a*) the approaches implemented by the the selected repositories, (*b*) the gap between the expectations and the current support, and (*c*) some proposals to fill the gap.

4.1 Dataset Formatting

From the perspectives of data publishing and data repositories, a dataset is a unit of information and there are at least two notions of format associated with it: (*i*) the *file format*, i.e., the way a dataset is encoded in one or more files; and (*ii*) the *content format*, i.e., the way a dataset content is actually organised. For instance, when dealing with tabular data the file format can be csv while the content format can be a description of the organisation of the data in the table. The notion of content format includes that of data file metaformat (Raymond, 2008; Parsons and Duerr, 2005) yet it is a bit more abstract, e.g., it does not imply that there is a standard software library for marshalling and unmarshaling it. Having data appropriately formatted is a pre-requisite for any use of it. The file format impacts on the capacity of current and future software to "import" the dataset content while the content format impacts on the interpretation and understanding of the dataset.

Repository practices. All the analysed repositories explicitly deal with files and their formats. Data owners are requested to provide the repositories with files realising the datasets and data consumers are provided with files when they access the dataset content. For content format, selected repositories somehow neglect it and describe how the dataset is organised through dataset documentation (cf. Sec. 4.2). Thus in the remainder, the analysis focuses on file format only.

All repositories in our sample accept datasets stored in files in any format, there is no particular restriction on allowed formats. However, data owners are often encouraged to submit data in "friendly" formats, namely standard formats that are supposed to be suitable for preservation and reuse. For example, 3TU. Datacentrum provides users with a table including the "preferred formats" for any type of file, namely the optimal file formats used for long-term preservation of data (cf. Sec. 4.6). In general, such formats are non-proprietary and well documented. Examples of formats recommended by Dryad include csv, XML and JSON as well as community standards (where they exist). The CSIRO DAP declares the ad-hoc formats supported for specific datasets. For example, all files composing the ATNF Astronomy Observations datasets are compliant with PSRFITS, i.e., a standard for Pulsar Data Storage (Hotan et al., 2014).

Detailed information on the most common file formats used by the published datasets is given in **Table 5**. Overall, it can be observed a large variety of formats including: (*a*) compressed archives thus making it not immediate to figure out how many files they actually contain and what are their formats, (*b*) generic formats ranging from pdf to tabular formats like csv and Excel, and (*c*) domain specific formats, e.g., in CSIRO there is a large use of formats that are common in astronomy[10].

Independently of file formats, repositories have some limitations on allowed file sizes. They tend to have an upper bound limit yet are open to negoziate extensions to this limit with additional costs (cf. Sec. 4.4). Dryad allows uploading of no more than 10GB of material for a single publication; 3TU.Datacentrum supports the upload of datasets up to 4 GB; Zenodo currently accepts files up to 2GB although it reports that

[9] This is the validation feature.

[10] The PSRFITS (Hotan et al., 2014) pulsar data format envisages ".sf" for raw data files – search mode data, ".rf" for raw data files – folded data, ".cf" for calibration data files, and ".FTp" for pulsar total intensity profile – averaged over time and frequency.

Format	3TU.Datacentrum	CSIRO	Dryad	Figshare	Zenodo
#1	app./x-netcdf (3070 – 80%)	app./fits – sf (143,376 – 26%)	text/plain – txt (4926 – 16%)	*n/a* – xls (267,222 – 83.3%)	*n/a* – sav (1798 – 15.2%)
#2	app./zip (559 – 14.6%)	image/png – png (95,072 – 17.3%)	Excel 2007 – xlsx (3793 – 12.3%)	*n/a* – pdf (16,996 – 5.3%)	*n/a* – txt (1243 – 10.5%)
#3	text/plain (57 – 1.5%)	app./fits – rf (94,028 – 17.1%)	text/csv – csv (3099 – 10%)	*n/a* (12,968 – 4%)	*n/a* – png (1059 – 9%)
#4	app./octet-stream (27 – 0.7%)	app./fits – FTp (92,888 – 16.9%)	app./zip – zip (2834 – 9.2%)	*n/a* – docx (4868 – 1.5%)	*n/a* – fits (1043 – 8.8%)
#5	app./x-hdf5 (22 – 0.6%)	app./fits – cf (82,204 – 14.9%)	Excel – xls (2074 – 6.7%)	*n/a* – doc (4511 – 1.4%)	*n/a* – zip (878 – 7.4%)
#6	app./x-gzip (19 – 0.5%)	*n/a* – adf (9833 – 1.8%)	*n/a* – *n/a* (2007 – 6.5%)	*n/a* – xlsx (4012 – 1.2%)	*n/a* – gz (616 – 5.2%)
#7	video/x-msvideo (10 – 0.3%)	*n/a* – dat (4920 – 0.9%)	text/plain – nex (1191 – 3.9%)	app./zip – zip (1988 – 0.6%)	*n/a* – csv (532 – 4.5%)
#8	video/mpeg (9 – 0.2%)	*n/a* – nit (4911 – 0.9%)	app./pdf – pdf (1097 – 3.6%)	*n/a* – csv (1422 – 0.4%)	*n/a* – csv (269 – 2.3%)
#9	app./x-gzip (8 – 0.2%)	image/tiff – tif (3926 – 0.7%)	app./x-gzip – gz (734 – 2.4%)	*n/a* – jpg (1395 – 0.4%)	*n/a* – itp (260 – 2.2%)
#10	app./zip (4 – 0.1%)	*n/a* – 001 (1637 – 0.3%)	app./x-fasta – fasta (728 – 2.4%)	*n/a* – cif (1108 – 0.3%)	*n/a* – ods (205 – 1.7%)
Distinct	53	1876	868	524	961

Table 5: Top 10 formats associated with published datasets. Format is mime type – file extension.

the current infrastructure has been tested with 10GB files; Figshare enables users to store up to 1 GB data in their private space with files up to 250 MB each.

What is missing and trends. As already outlined, generalist repositories cannot make any assumptions on the data file and content formats they are requested to deal with. This has lead to the development of approaches that aim to be generic and as much format agnostic as possible (see also Sec. 4.6). This situation largely restricts the facilities that a repository can offer to support data publication. Visualization, for example, is not always guaranteed. Thus, it is very important to make scientists depositing their datasets fully aware of what services a repository can or cannot offer upon the upload of a dataset in a certain file format. It is expected that making this warning explicit can contribute to progressively foster the usage of a more restricted set of (de-facto) standards on which a richer pool of services for facilitating the exploitation of published data can be made available.

Another key trend concerns identifying and promoting dataset formats that are as much 'intelligible' as possible to permit cross disciplinary (re)use of datasets. An example of ongoing efforts in this direction is given by The Open Grid Forum that is working on the definition of a Data Format Description Language (Beckerle and Hanson, 2014) for describing text, dense binary and legacy data formats in a declarative manner. Another example is provided by the Research Data Alliance (RDA) Data Type Registries Working Group (Broeder and Lannom, 2014) that is discussing how to enable associating an intelligible 'type' to a dataset. The notion of 'type' promoted by this WG is aiming at characterising the dataset at multiple levels of granularity, i.e., from individual data points up to the entire dataset. They propose to use a federation of registries to make it possible for data consumers to discover an accurate description of a given type as well as any additional information for managing such a type, e.g., potential software for processing data of the given type. However, this approach has to face the potential limitations that have been widely discussed in the past, e.g., McGath (2013) stressed how new formats develop and existing ones evolve officially and unofficially thus making the maintenance of the registry challenging. The availability of this meta-registry (or federation of registries) would also facilitate repositories willing to enlarge the set of formats a dataset can be accessed. Actually, the set of formats a dataset is made available can either belong to a pre-defined list or be the result of a negotiation between the data consumer and the repository itself aiming at identifying the format that better fits the purpose of the consumer (Parsons and Duerr, 2005).

4.2 Dataset Documentation

The ultimate goal of data publishing is to make datasets available for validation and reuse both within the scientific community that has produced it and, more widely, within other communities. In order to fully meet this objective a dataset produced by a community of practice has to be endowed with auxiliary data providing contextual information about the dataset, like what it is about – descriptive metadata – and how it has been obtained – data provenance (Simmhan et al., 2005; Carata et al., 2014). This documentation at large, deeply influences dataset discoverability, understandability, verification and practical re-use (Thanos, 2014).

Various approaches are adopted for equipping a dataset with proper documentation. These include the annotation of the dataset with appropriate metadata maintained in, and made available through, the dataset repository itself as well as the publication of a specific data paper about the dataset (Candela et al., 2015) containing, among the other information, a link to where the dataset has been deposited. However, producing effective and exhaustive documentation is really challenging because of the open-ended re-use scenarios. It is almost impossible to anticipate all (re-)use of a dataset and thus identify a suitable designated community (Parsons and Duerr, 2005; Palmer et al., 2011).

Repository practices. All the analysed repositories require that basic metadata are specified when submitting a dataset; none of them explicitly requires that the deposited datasets be associated with a data paper.

In analysing the publicly available documentation we could not find out any specification of which metadata are internally supported by the selected repositories. In order to identify such metadata, we then analysed, for each repository, the metadata requested at submission time and the metadata exposed at visualisation time, i.e., the metadata returned when a repository user access the dataset landing page. We found that eleven classes of attributes are used:

- *Availability*: enabling to get access to the dataset and its content, e.g., a DOI or a URI;
- *Bibliometric data*: reporting dataset statistics including number of data visualizations and downloads;
- *Coverage*: describing dataset "extension", including spatial, temporal and taxonomic coverage;
- *Date*: providing information about "when" the dataset was created, submitted and published, including embargo period;
- *Format*: characterising the dataset from the formatting perspective (namely file format) including the size;
- *License*: describing the policies ruling the dataset reuse, including access rights and licenses;
- *Minimal description*: analogous to basic scholarly publication ones such as title, authors, brief description or abstract;
- *Paper reference*: providing reference(s) to related publication(s), including a DOI or a URL;
- *Project*: describing the initiative leading to the production of the dataset, including research goals, type of research and funding sources;
- *Provenance*: specifying the methodologies leading to the production of the dataset, including original sources, instruments and software tools used to create the data files;
- *Subjects*: including keywords, classification codes, tags, and subject headings.

A summarising picture of the classes of metadata attributes actually supported by each repository is given in **Table 6**. Repositories make different choices implementing these classes. There is a large heterogeneity with respect to how many attributes repositories support per class, whether an attribute is mandatory or not, whether attribute values are compiled by using controlled vocabularies or free text. For example, all the selected repositories envisage paper reference information but only for Dryad this information is mandatory. Dryad also requires very specific attributes about articles associated with the dataset, including title, authors, journal, abstract, keywords and coverage; instead, Zenodo only suggests to provide information about journal name, volume, issue and pages.

At the moment of submission, repositories may ask submitters to upload other types of supporting documentation. This is the case of CSIRO DAP and Dryad that encourage authors to provide additional documentation in the form of "ReadMe" files for helping proper interpretation and reuse of the dataset. In particular, Dryad recommends the ReadMe be a plain text file containing the following information: (*a*) for each file of the dataset, a short description of which datasets are included; (*b*) for tabular data, definitions of column headings, row labels, data codes and measurement units; (*c*) any data processing steps that may affect

	3TU.Dat.	CSIRO	Dryad	Figshare	Zenodo
Availability	✓	✓	✓	✓	✓
Bibliometric data	✓		✓	✓	
Coverage	✓	✓	✓		
Date	✓	✓	✓	✓	✓
Format	✓		✓	✓	
License		✓	✓	✓	✓
Minimal description	✓	✓	✓	✓	✓
Paper reference	✓		✓	✓	✓
Project	✓	✓			✓
Provenance		✓			
Subjects	✓	✓	✓	✓	✓

Table 6: Dataset attributes supported by Scientific Data Repositories.

interpretation of results; (*d*) a description of the associated datasets that are stored elsewhere, if any; and (*e*) contact information for questions.

What is missing and trends. It is now becoming evident that the use of purpose-oriented metadata is of paramount importance to achieve reusability. However, the current metadata descriptions supported by generalist repositories are necessarily limited in this respect since these repositories are meant to serve multiple unspecified target communities that may access published data for unconstrained re-uses. Previous studies on metadata for scientific data (Willis et al., 2012) highlighted that (*a*) there is a large variety of metadata schemas in use (over 50 were reported), (*b*) there are commonalities in existing schemas across disciplines and types of data, and (*c*) there is the need for more metadata-related research to actually increase the access to and reuse of research data. To overcome the limitations that necessarily result from the use of generic metadata specifications, data owners are starting to associate a data paper with the dataset. A data paper (Candela et al., 2015), if well written, can provide a data consumer with the details facilitating understanding and potential reuse of the data. Unfortunately, a data paper is of use only for humans and does not provide any support for automatic consumption. The same comment applies to ReadMe-based documentation.

A contribution to better discoverability and reusability might be obtained by enabling multiple metadata descriptions and documentation for the same datasets, each oriented to the needs of a specific target audience for a particular application (Parsons et al., 2011). These descriptions might be either provided by the data owner or partially generated automatically. Similarly, multiple data papers might be prepared each providing a focussed presentation to be published in a journal addressing the corresponding specific domain. Data users might select, among the multiple alternatives, the metadata, the data paper or other documentation that best meet their needs. This selection might be facilitated by the presence of Metadata Standards Directories, whose introduction as global infrastructure resources is currently being largely discussed (Ball et al., 2014). If appropriately developed these directories might support an automatic treatment of the information about metadata formats that might be exploited not only by humans but also by third-party services exploiting the described dataset.

At the moment metadata associated with datasets are primarily meant to provide a static description of the data, i.e., when, who, where and how they have been gathered or produced. However, it is well known that the knowledge about certain data may increase over the time. In particular, data validation and reuse activities may contribute to refine the understanding of the data potential and their limitations. For the sake of re-usability, it is very useful to be able to add further metadata values and documentation as the data exploitation progresses. These additional information might, for example, provide more refined information on the data fit-for-purpose. This approach resembles the mix of "publishing" and "push" models suggested by Kansa et al. (2014) where various actors collaborate in "producing" quality datasets. In the case of the classes of attributes observed in the analyzed repositories only "Bibliometric Data" and "Paper Reference" may evolve over the time. Bibliometric data – reporting statistics on the dataset usages – is indeed useful as an indicator of trust in the dataset. Paper reference provides a means for more in-depth understanding of

the dataset as it links papers describing results and experiences, both positive and negative ones, obtained by exploiting the given data (cf. Sec. 4.5).

Enriching dataset metadata with links to related papers is just an example of an important trend that is growing the last years. Linking is progressively introduced to add more contextual information to research outcomes, and thus to facilitate understanding and re-using them. Technological advancements, especially in mining techniques, data availability, and development of data infrastructures offering specialised services on large amount of content are today largely facilitating the automatic creation of such links. For example, experiments to automatically create links between datasets and the papers citing these datasets (Manghi and Mannocci, 2013) have recently been performed by exploiting OpenAIRE infrastructure services.

4.3 Dataset Licensing

A licence is a legal instrument for a dataset owner to state the terms of use of the published dataset by third parties (Ball, 2012). A default legal position on how the dataset may be used in any given context is hard to be derived if not explicitly specified (Campbell, 2015; Guibault and Wiebe, 2014). It involves many factors related to the dataset context and may also depend on external parameters, e.g., jurisdictions.

When publishing occurs through a repository service there are two types of licenses involved: (*i*) the one agreed between the repository and the data owner and (*ii*) the one agreed between the repository and the data consumer. Both these licences are partially captured by the "terms of services" or "policies" of the repository, i.e., they are part of the rules a repository user must agree to accept when using the repository service. Repositories usually support a single deposition license that specifies the rights expected to be granted to them by the dataset owner, e.g., for curatorial or usage tracking functions. They are commonly more flexible for what concerns dataset re-use supported licences. A re-use license concerns at least attribution, copyleft requirement and control on commercial exploitation of the dataset. One of the first steps that a data owner is requested to do when depositing a dataset in a repository is thus to check whether the licence he/she is willing to associate with the dataset is compatible with one of those supported by the repository. Concerning the type of licenses there is nowadays a progressive orientation, especially in the context of the publicly funded research, to move from the usage of proprietary licenses in favour of established ones, e.g., the Creative Commons licenses (https://creativecommons.org/). This simplifies specification, understanding and communication of such licenses.

Repository practices. For deposition licences, data owners are requested to register in order to use the repository facility. Once registered, whenever data owners submit a dataset they are implicitly accepting the repository policies. These policies are quite heterogeneous:

- *3TU.Datacentrum*: data owners grant the repository a non-exclusive license for the digital data files enabling repositories to store the dataset and make them available to third parties;
- *CSIRO*: data owners grant the repository the right to make the datasets available on a non-exclusive, non commercial basis yet accompanying the dataset with terms of use at download time;
- *Dryad*: data owners grant the repository permission to make the dataset available to the public under a CC0 waiver (although other licensing terms may be permitted);
- *Figshare*: data owners agree that any publicly stored material (including datasets) will be made available with Creative Commons Licenses (CC-BY for filesets and CC0 for datasets);
- *Zenodo*: data owners must specify the license governing dataset content by choosing from a wide variety of available licences.

In all these cases it is possible to specify an embargo period, i.e., a period of time where access to dataset content is restricted. In addition to embargo periods, some repositories (e.g., Figshare and Zenodo) make it possible to upload datasets to be kept "private" (this use of the repository facility can be seen as a sort of pre-publication phase where the dataset owner moves the data from its premises to the repository premises). In no case the upload of datasets lead to changes of ownership.

For access licences, in practice repositories are expecting to serve the general public thus they do not require data consumers to register. Publicly available datasets can be discovered and accessed by every user (cf. Sec. 4.7) without the need for authorisation or login. It is worth to highlight that granting access or restricting access to dataset content is the only access licence repositories can enforce in practice.

In addition to access licences, repositories deal with datasets terms of use, i.e., the licence characterising how the dataset content can be actually used once successfully accessed. The 3TU.Datacentrum and CSIRO DAP repositories have defined their own proprietary licenses that users have to comply with.

In particular, the general conditions of use established by 3TU.Datacentrum resemble the Creative Commons Attribution-NonCommercial License (CC BY-NC), stating that any user wishing to reuse a dataset stored in the repository must always acknowledge the dataset sources and in any case the dataset must not be used for commercial purposes. Similarly, the CSIRO Data License grants user a royalty-free, nonexclusive, non-transferable license for using the dataset only for noncommercial purposes. In the case of Dryad, Figshare, and Zenodo the default option is CC0. However, Figshare uses CC-BY for datasets stored as filesets while Zenodo makes it possible for data owners to associate the licence they prefer by choosing from a list of existing ones. Detailed information on the terms of use licences actually associated with the published datasets is given in **Table 7**. From this analysis it emerges that the number of actually used licences is quite limited. Moreover, the use of Creative Commons licences is very diffuse while the use of proprietary or "in house" licences is very limited.

Terms of use licences are part of the documentation repositories associate with datasets (cf. Sec. 4.2). As expected, the good practice of having the terms of use licences clearly reported on the dataset landing page is quite diffused. In the case of Dryad, this is actually among the very first information about a dataset. In the majority of cases, the acceptance of the terms of use is tacit, it is part of the acceptance of the terms of use of the repository service. In the case of CSIRO, the data consumer is explicitly requested to accept the terms of use before being able to actually download the dataset content.

To conclude the discussion on licences, it is worth to highlight that repositories have no real instruments to enforce dataset terms of use, e.g., they can not guarantee that dataset consumers act according to the given licences.

What is missing and trends. Dataset licensing impact on various aspects of dataset re-use ranging from attribution to commercial exploitation. Repositories should reinforce the support offered by data owner at deposition time by clearly reporting the impact diverse licences have on the dataset usages, e.g., by clearly indicating the difference induced by a CC0 licence with respect to a CC-BY licence on attribution. They should offer an easy to use dashboard to compare licences, e.g., Daga et al. (2015).

As discussed above, generalist repository approaches are limited in scope with respect to dataset licensing and this makes it difficult the effective re-use of data, out of the most trivial cases. This is indeed a broad and challenging issue that still requires considerable work for identifying explicit and standard solutions being able to avoid "interoperability" issues. Eschenfelder and Johnson (2014) have recently conducted a detailed analysis of access and use control policies in a subset of repositories they call "controlled data collection", i.e., repositories where staff, or user communities, make and enforce rules to control who can access data or how data can be used. In their conclusions, they state that repositories should offer a vast array of use and control options, as "one-size-fits-all solutions" do not exist. Managing and applying some forms of control on access and, especially, re-use conditions in heterogeneous contexts as those generalist repositories are confronting with is one of the most relevant yet unsolved issues. Despite few languages to specify rights exist, e.g., MPEG-21 REL (Wang et al., 2005) and Open Digital Rights Language (Iannella, 2002), meeting the aim to formally

Licence	3TU.Dat.	CSIRO	Dryad	Figshare	Zenodo
#1	*n/a* (3453 – 97.85%)	CC-BY 3.0 (870 – 61.35%)	CC0 1.0 (29,025 – 94.19%)	*n/a* (308,108 – 96.16%)	CC0 1.0 (1041 – 64.7%)
#2	CC BY-SA 3.0 (22 – 0.62%)	CSIRO Data Licence (328 – 23.13%)	*n/a* (1745 – 5.66%)	CC-BY (12,262 – 3.83%)	CC BY 4.0 (251 – 15.6%)
#3	(c) CITG Delft (18 – 0.51%)	CC-BY 4.0 (83 – 5.85%)	Unknown (8 – 0.02%)	CC0 (41 – 0.01%)	openAccess (175 – 10.88%)
#4	Delft, KWR (16 – 0.45%)	No Licence (47 – 3.31%)	Custom (1 – *n/a*)	Apache-2.0 (2 – *n/a*)	CC BY-SA 4.0 (72 – 4.47%)
#5	Public (12 – 0.34%)	CC BY-NC-ND 3.0 (45 – 3.17%)	Custom (1 – *n/a*)	GPL-3.0 (1 – *n/a*)	closedAccess (50 – 3.11%)
Distinct	10	10	39	6	7

Table 7: Top 5 licences associated with published dataset. '(c) CiTG Delft' is 'Delft University of Technology, Civil Engineering and Geosciences'; 'Delft, KWR' is 'Delft University of Technology, KWR Watercycle Research Institute'; 'openAccess' is 'info:eurepo/semantics/openAccess'; 'closedAccess' is 'info:eu-repo/semantics/closedAccess'.

specify rights is unrealistic unless specific controlled vocabularies are introduced and agreed. The Creative Commons Rights Expression Language, for example, goes in this direction by introducing precise definition of its terms, as expressed by the Creative Commons licences (Abelson et al., 2008).

Another important aspect that strongly influences the possibility of effectively reusing published dataset, which has not yet received enough attention, is "compatibility" between the licenses of the dataset and the licences of the rest of material "associated" with the dataset, including its documentation. The capability to actually reuse the dataset depends on the capability to use the surrounding information and tools characterising the dataset. These elements should themselves be published with clear licences that do not invalidate access and reuse of what is indeed needed to complement the understanding and use of the data. In particular, the ability to manage and control compatibility would provide the base for removing the difficulty often mentioned by the scientists when they have to decide whether providing an open or restricted access to their datasets. By exploiting this facility the repository might support the publishing of different versions of the same dataset each characterised by a diverse licence thus facilitating data sharing to the maximum extent it is possible. Flexible and powerful mechanisms enabling to control dataset access and reuse are expected to reduce the tendency to avoid the publishing of potentially "problematic" datasets, e.g., datasets with privacy issues (Kowalczyk and Shankar, 2011).

4.4 Dataset Publication Costs

The cost is one of the major factors usually hindering data publishing to be a norm in science. It includes the effort needed to prepare the datasets in a way that enable others to make use of them, e.g., to format and document the dataset, as well as the monetary cost for having the dataset archived in a trustworthy repository that also provides access to the dataset content.

Our analysis of the data publication costs is focussed on the monetary cost, if any, dataset owners and providers are exposed to when using the repository service. Other forms of cost related with the use of the service – e.g., time needed to "familiarise with" and "use" the service – or the cost repository managers have to sustain in order to offer their service are not discussed.

Repository practices. Operating a repository service has a cost. The tendency of all the selected repositories is to charge data owners when publishing their dataset rather than data consumers.

The Dryad cost model largely resembles the open access model adopted by many journals, its datasets are mainly expected to be openly available to the public (cf. Sec. 4.3), but a charge is always required at the time of data submission. In particular, the repository requires the submitter to pay a Data Publishing Charge (DPC), unless (*a*) "the submitter is based in a country classified by the World Bank as a low-income or lower-middle-income economy", or (*b*) "the journal publishing the research article associated with the dataset has already contracted with Dryad to cover the DPC". In order to encourage organisations to cover the DPCs on behalf of their researchers, Dryad offers a series of plans providing for volume discount ranging from a voucher plan, e.g., pay for the publishing of a number of data packages, to subscription plans, e.g., pay an annual fee. The cost for a single data package up to 10GB is $80. For data packages exceeding the 10GB size limit, $15 are charged for the first GB and $10 for each additional GB or part thereof. Moreover, for journals that do not use the integrated data submission service offered by the repository, the submitter has to pay an additional $10 fee at the time of submission to cover added curation costs. Dryad is the only repository in our sample that always requires submitters to pay a charge independently of the files size. All the other ones offer at least a minimum storage space where users can publish free of charge.

Figshare includes a completely free plan, which allows users to store private data up to 1 GB at no charge, with a size constraint of 250 MB per file. Payment only involves private storage, as the space for publishing public data is always unlimited for every plan. The other plans require users to pay a monthly amount depending on the dimension of the private storage space and the file size limit, e.g., a fee of 8$/month guarantees a private space of 10 GB with a size constraint of 500 MB per file.

Zenodo offers its service for free yet it currently imposes a size constraint of 2 GB per file. However, it does not want to exclude larger datasets, so it is planning to put a ceiling to the space that can be offered at no charge, and to introduce payment plans for bigger data.

3TU.Datacentrum requires no payment for datasets up to 4 GB. For larger datasets, users have to contact the repository's staff for arranging a customized upload.

What is missing and trends. Incentives and mandates aiming at enlarging the amount of published data have limited impact if they are not accompanied by measures addressing the reduction of the publishing costs, so that researchers are not discouraged from publishing their datasets (Roche et al., 2013). This could

be achieved, for example, by explicitly relating costs with offered services and their quality. Not all the datasets, indeed, require the same level of curation or preservation for the same number of years, or the same level of quality of services in accessing them. Clearly, the quality of these services largely influence the publishing costs. The analysed repositories often do not specify in detail the quality of service offered. Thus for those that decide to publish their datasets it is hard to understand the motivations behind the costs and to select the solution that best satisfy their publishing needs. A more detailed specification of the offered services would also facilitate a fair distribution of the costs among the actors involved in the data publishing process. Actually, the cost of the repository services has not necessarily to be entirely covered by the researchers. This expectation is strictly tied to the strategies and efforts that are presently put in actions to promote "open access" to research data, strategies and efforts that are solicited by the overall recognition of the economic benefits so produced (OECD, 2007; Nicol et al., 2013). Rather than billing researchers, new payment methods can be envisaged including private- or public-sector grants and partnerships with journal publishers. For instance, Figshare has already started partnerships with some Publishers to support authors who wish to openly share their datasets and articles' supplemental material, e.g., PLOS (figshare, 2013), Taylor & Francis Publishers (Devine, 2014), John Wiley & Sons (Peters, 2015). A lot can still be done to reduce the repository operational cost. Current trends in this direction are mainly devoted to rely on specialised third-party providers for well defined functionality, e.g., acquiring the needed resources from cloud providers rather then spending effort in operating such resources on their own or outsourcing preservation to specialised centers. There also are many attempts to exploit techniques that reduce the costs of curation by automatically extracting part of the necessary information through analysis and mining of related artefact, e.g., papers. With the progress of interconnected data source infrastructures this is a very promising area of development.

4.5 Dataset Validation

Dataset validation is an essential phase of the data publishing endeavour since it is expected to somehow contribute to assess the "quality" of the dataset. Its real meaning is at the moment among the most debated and undefined data publishing phases. It refers to any process aiming at assessing the "cogency" or "soundness" of the published data. Very often this validation is intended to share commonalities and objectives with the peer review process applied in literature (Lawrence et al., 2011; Mayernik et al., 2014). However, differently from the peer-review of papers, there are no shared established criteria on how to conduct such review and on what data quality is. Actually, it has been observed that the use of the term "peer review" for datasets is causing some false expectations (Parsons and Fox, 2013; Candela et al., 2015).

Scientific data repositories help the dataset validation phase by offering practices and services for both dataset *pre-publication* and *post-publication* validation. The former is a sort of preventive assessment aiming at avoiding that "poor quality" datasets are published, it is a sort of quality assurance involving both the data owner and the repository acting as data publisher. The latter is a validation resulting from feedback received from users trying to actually reuse the published dataset, be it positive or negative.

Repository practices. The analysed repositories support the pre-publication phase by using techniques that aim at verifying that datasets are not corrupted and that their associated metadata is syntactically correct and complete. No control is made for attesting "scientific soundness" of the dataset, e.g., scientific validity, accuracy or completeness.

All the repositories in our sample, with the sole exception of Figshare, perform a validation process of both data and metadata. However, only Dryad gives accurate information about this process in its 'Terms of Service'. In particular, Dryad curation personnel performs a series of checks ranging from technical ones (e.g., files can be opened, are not corrupted, do not contain viruses) to administrative ones (e.g., metadata is technically correct, information on the associated paper is in place). Dryad may review the content for reasons including the presence of inappropriate information and copyright statements incompatible with CC0. In any case, Dryad does not check the dataset from the scientific perspective or modify the content except for accessibility reasons. Besides these pre-publication controls, repositories perform checks aiming at guaranteeing that datasets content remains intact. 3TU.Datacentrum, Dryad and Zenodo store all datasets along with a checksum of their content. For example, data files submitted to Zenodo are stored with a MD5 checksum of their content, and regular checks of files against their checksums are made.

For post-publication validation, repositories tend to offer top counts or statistics on datasets download or usage. Figshare offers the possibility to know how many times each dataset has been shared or viewed through its search or browse option. In the first case, information on views and shares are given with the

dataset description. With the browse option, it is possible to select a given category from a menu and then produce an ordered list of all the datasets classified in that category. The produced list can be sorted by "most shared" or "most viewed". A similar approach is offered by Dryad that records the number of downloads and makes it possible to browse the "most popular" ones, in terms of downloads. 3TU.Datacentrum only publishes aggregated statistics on downloaded datasets as a sort of validation of the centrum itself.

What is missing and trends. Dataset validation is a complex process still far from being completely characterized. There are many reasons for this complexity including: (*a*) it is a shared responsibility – there is no single actor in the scientific data publishing scenario that can take the responsibility of assessing the validity of a dataset; (*b*) it is a continuous process – usage of a published datasets is not confined to a given community or to a use case only, any usage of a published dataset may potentially require a validation assessment; (*c*) it is a many facets issue – it has to do with technical, scientific, and organisational aspects (to cite a few) all requiring diverse domain expertise; (*d*) it is not a matter of repository certification – it is almost impossible to envisage a certification scheme that guarantees that the datasets published by a repository are "scientifically sound".

In the context of generalist data repositories dataset validation is at its initial stage. It is characterized by attempts to define what "data quality", validation criteria and reviewing is in general terms, i.e., without any assumption on the domains where the datasets have been produced and will be consumed. Why validation practices are so scarcely applied by data repositories is discussed in a recent blog post (Kratz, 2014). Among the suggestions given in this blog, there is "forget quality, consider fitness for purpose". The motivation for this is that "A dataset may be good enough for one purpose but not another". Trying to assess the general 'quality' of a dataset is hopeless; consider instead whether the dataset is suited to a particular use. Extending the previous idea, documentation of how and in what contexts a dataset has been used may be more informative than an assessment of abstract quality".

Repositories might help realizing the suggestions expressed in Kratz blog by favouring the production of documentation (cf Sec. 4.2) that is sufficiently rich to assess whether the dataset fits for users' purpose. For example, they might implement pre-publication validation procedures and resources aiming at assessing the published artefact (both the dataset and the documentation) by using an "average consumer" perspective. This perspective consists in analysing the artefact without any domain specific knowledge about the dataset. Repositories might also support post-publication validation by providing mechanisms for end-users to provide both the repository and the data providers with concrete and documented feedback resulting from (re-)using or attempting to (re-)use the dataset. This feedback may contribute to authors' scholarly records and may also have the traits of a scientific publication. The authors of the feedback (data consumers) might give arguments on their experience in using the dataset including the scientific questions they are posing, their application domain, and the competing interests. This feedback would importantly contribute to enrich the documentation accompanying a published dataset. In addition to this explicit and feature-rich feedback, repository services might also heavily exploit the advances resulting from the altmetrics research aiming at collecting diverse "flavours" of impact of scholarly outputs including datasets (Costas et al., 2013; Andreoli-Versbach and Mueller-Langer, 2014).

4.6 Dataset Availability

Dataset availability is the feature aiming at guaranteeing that published datasets are at consumer's disposal. It is among the key features a scientific data repository is called to support in the dataset publishing settings. Repositories offer mechanisms for *present availability*, i.e., datasets are available at access time (cf. Sec. 4.7), and *future availability*, i.e., datasets are available over time.

Repository practices. Two main approaches are implemented to guarantee present and future availability, i.e., archiving data in a secure manner and apply data preservation approaches.

For secure archiving of data, all the analysed repositories use to store multiple copies of the datasets either on their own premises or on third party service providers, e.g., Figshare uses Amazon facilities, Zenodo uses the CERN Data Centre. In addition, both Dryad and Figshare have partnered with the CLOCKSS organisation, a geographically and geopolitically distributed network of 12 redundant archive nodes located at 12 major research libraries around the world. A failover copy of all contents published by Dryad is maintained in the CLOCKSS Archive, so, if Dryad could no longer maintain the repository as an active service, all Dryad-registered DOIs would be updated to resolve to the copy at the CLOCKSS archive, which would continue to provide access to the content under the same licensing terms. A similar solution is implemented by Figshare.

For preservation, beside storing the data in multiple copies the selected repositories tend to use format migration practices. Format migration is a challenging feature to be guaranteed due to the almost open ended set of data formats (file formats) that repositories are called to manage (cf. Sec. 4.1). In fact, some repositories explicitly declare that they do not guarantee usability and understandability of deposited objects over time, e.g., Zenodo and Dryad. However, Dryad announces to perform format migration, i.e., a new version of a file in a migrated format may be created and added to the original dataset whenever the repository judges that this may facilitate preservation. Migrated files may not contain all the information available in the original file format, but the repository tries to minimize information loss caused by file format migration. In any case, the information content of the original file is never modified. 3TU.Datacentrum highly stresses the importance of taking appropriate measures to guarantee future accessibility to the data. The License Agreement between 3TU.Datacentrum and data submitters explicitly states that the repository (*i*) "shall ensure, to the best of its ability and resources, that the deposited dataset will remain legible and accessible"; (*ii*) "shall, as far as possible, preserve the dataset unchanged in its original format, taking account of current technology and the costs of implementation"; and (*iii*) "has the right to modify the format of the dataset if this is necessary in order to facilitate the digital sustainability, distribution or re-use of the dataset". 3TU.Datacentrum provides users with a table including the optimal file formats used for long-term preservation of data. The number of supported formats is limited, in order to facilitate future conversion to other formats, but the table is regularly updated with new formats. For each format, the level of support for long-term preservation is indicated according to three levels: Level 1 – "All reasonable actions to maintain usability will be taken; actions may include format migration, normalization or conversion"; Level 2 – "Limited steps to maintain usability will be taken; file formats may be actively transformed from one format to another to mitigate format obsolescence", and Level 3 – "Only access to the object in its submission file format is provided".

What is missing and trends. In order to guarantee current and future availability of published datasets repositories have to address both technical and financial challenges. From the technical perspective, long-term preservation of digital information is not a new concern, e.g., Berman (2008); Burda and Teuteberg (2013). However, dataset preservation in the context of generalist repositories has specific challenges mainly resulting from the almost open ended set of dataset typologies to be managed (cf. Sec. 4.1). Solving these challenges has necessarily a financial impact on repositories and the level of completeness of the proposed solutions depends on the availability of known approaches. For instance, if datasets are made available via formats requiring rare or proprietary programs to be interpreted, then implementing mechanisms for migration may be very expensive. Adopting a specific availability policy is thus necessarily a trade off between either an ideal or a feasible solution compatible with the resources at hand (including the financial ones).

Repositories can then decide to make different choices and offer different levels of availability. This level may change during the lifetime of the repository because of factors ranging from repository technological advances to capacity development. Unfortunately, today the availability policy implemented by a repository is rarely explicitly described. This makes it hard for users, be they dataset providers or consumers, to select the policy that best fits their needs. In order to better support publishing needs repositories should then not only design and develop their availability policies but also make them explicit with the goal to prevent false expectations. The availability policy should be an integral part of the Terms of Service every scientific data repository should have. In this regard the approach used by Dryad to develop its preservation policy and described by Mannheimer et al. (2014) is a good example.

Availability is usually related also with repository trustworthiness, e.g., Kratz and Strasser (2014). Even if there exist certification schemes for assessing this trustworthiness, e.g., Dobratz and Scholze (2006), very few scientific data repositories are "formally certified". Thus trustworthiness is actually an induced property acquired by repositories during their lifetime.

4.7 Dataset Discovery and Access

Dataset discovery and access is the facility enabling consumers to become aware of the existence of certain datasets and be able to get access to them, namely to the dataset content (its files) and the associated documentation. It is the basic facility scientific data repositories are called to offer to data consumers in the context of data publishing. This facility might include user-driven functions, e.g., search, browse, click to download, as well as semi-automatic functions, e.g., notifications and recommendations. Usually, these facilities strongly rely on dataset documentation (cf. Sec. 4.2), generally metadata.

Repository practices. The discovery facilities offered by the repositories analysed in this survey are summarised in **Table 8**. All of them offer the following well known approaches for datasets discovery:

- *Keyword-based search* Users can specify their information need through a set of keywords;
- *Field-based search* Users can specify their information need through a set of field-based filtering criteria. The set of supported fields is repository specific;
- *Browse* Users are provided with a list of datasets to scan. This list can be either the entire list of datasets published by the repository or the list of datasets resulting from a given selection criterion, e.g., keyword, subject, type, format, year, creator. Criteria are repository specific.

In the table a distinction is to be made among those repositories offering browsing through a classification system and those that do not. In fact the first ones, namely CSIRO and Figshare, not only allow users to browse the datasets classified under the research field they are interested in, but also give them the capability to view details of dataset description and be able to express further keyword searches based on terms assuring positive results.

CSIRO offers two specific searches also: (*a*) search by location, i.e., users can specify their information needs by using a map to indicate the area of their interest plus keywords for narrowing the search, and (*b*) collection specific search, i.e., users are provided with forms enabling them to specify precise information needs on specific datasets including pulsar observations, microscopy images and sensor data.

3TU allows browsing by location to identify datasets for which geolocation is essential.

For access, all the repositories in our sample offer the possibility to download a dataset as a whole, as well as to download single files one by one, through apposite links displayed on the dataset web page. For example, CSIRO DAP displays every dataset as a file structure, with checkboxes for selecting specific files to download.

Apart from the basic download facilities, repositories provide alternative and customized modalities for accessing the datasets. For example, CSIRO provides access to datasets via protocols like WebDAV and SFTP. CSIRO also allows registered users to mount the data on selected machines and access the data directly on such machines. In some cases, users are provided with a "preview" of the dataset, e.g., CSIRO makes it possible to browse the images contained in a dataset, Figshare offers a preview of the dataset content in the browser.

Besides these human-oriented facilities, repositories support programmatic access to their content. This is achieved by supporting standard protocols, e.g., OAI-PMH (Lagoze and Van de Sompel, 2001), as well as web-based proprietary APIs. Such facilities are at the core of aggregative infrastructures (Manghi et al., 2014), e.g., OpenAIRE counts on the Zenodo OAI-PMH service to collect datasets published by this repository. In the case of programmatic APIs, they can also support the deposition phase, thus favouring the integration of the repository in publishing workflows.

What is missing and trends. The analysed data repositories implement data discovery facilities that are very similar to traditional ones offered by repositories of articles since many years. The criteria that researchers would like to formulate when looking for datasets, however, are often quite different from the ones they use for papers. Datasets are mainly searched to perform validation and reuse in scientific activities. To support these activities discovery criteria may require that datasets can be analysed by a given tool or that have been

End-user Facilities					
	3TU.Dat.	**CSIRO**	**Dryad**	**Figshare**	**Zenodo**
Keyword-based	✓	✓	✓	✓	✓
Field-based	✓	✓	✓	✓	✓
Browse	✓	✓	✓	✓	✓
Other	✓	✓			
Web-based API and Protocols					
	3TU.Dat.	**CSIRO**	**Dryad**	**Figshare**	**Zenodo**
Harvesting	OAI-PMH	OAI-PMH	OAI-PMH	In-house	OAI-PMH
Search	n/a	In-house	n/a	In-house	n/a

Table 8: Dataset discovery facilities. The CSIRO OAI-PMH facility is actually offered via the Research Data Australia service.

(re)used in research contexts "similar" to that of a given dataset, or even that have been approved or (re)used by coworkers or practitioners. Other criteria may be related to the contents of the dataset. For example, users may want to discover a dataset related to a specific geographic area or a dataset that better approximate it, maybe the only indication they have to discover the dataset is the title of a paper that reports results obtained by exploiting the dataset, or the link to an higher resolution version of the dataset that is not open as the one they are looking for.

A deep rethinking of repository discovery services that takes into account the specificity of the dataset resources and their usages is indeed necessary. Addressing this problem in a generic context is quite challenging. It requires an in depth analysis of the most common discovery patterns across domains and feasibility studies to understand to what extent the current quality of the metadata can support an effective discovery based on the new required paradigms. In order to enhance quality and amount of metadata and documentation associated with datasets, repositories might put in place solutions successfully applied in other contexts for exploiting contribution from third parties other than dataset providers. For example, tagging and rating could provide useful additional information to be usefully exploited at discovery time. On how to empower the set of dataset discovery facilities, it is just a matter of analysing the plethora of existing facilities aiming at making it possible for users to "discover resources" on the web and apply them to the datasets case. These facilities go well beyond the case of Google-like search engines. For instance, it is possible to envisage recommender systems for datasets (Bobadilla et al., 2013).

Discovery is a service that can also be provided by third-party providers, e.g., aggregative infrastructures and metasearch engines. In these scenarios protocols and APIs for collecting content (actually metadata) or collecting search results from many repositories play a key role. Repositories tend to support OAI-PMH (cf. **Tab. 8**) yet they should reinforce the range of protocols and facilities they offer for programmatically accessing their content, e.g., by exposing their content in formats other than HTML like Schema.org (Guha et al., 2016) and Linked Data (Bizer et al., 2009), by supporting standard protocols like OpenSearch and SRU (Denenberg, 2009). Services implemented by third-party providers largely rely on metadata, thus the quality of metadata severely impacts also the implementation of this facility e.g., Rousidis et al. (2014).

For data access, analysed repositories are actually relying on landing pages enabling to download files. It will be very useful if repositories will start supporting web-based protocols for accessing published datasets or part of them, e.g., (Cornillon et al., 2003). Although there is no protocol that is suitable for any dataset and access scenario, generalist repositories will enlarge their clientele by starting incrementally supporting existing protocols on selected dataset typologies.

4.8 Dataset Citation

Data citation is the practice of providing a reference to dataset(s) intended as a description of data properties that enable credits and attribution, discover, interlinking, and access to. It is a key mechanism in research data publishing since it enables data owners to get proper recognition for publishing their datasets and data consumers to explicitly refer to the datasets they are (re)used in their research. Scientific data repositories are called to play their role in supporting both owners and consumers.

Repository practices. The repositories in our sample offer various options to address data citation. These are summarized in **Table 9**. In particular:

- *Citation string* Allows users to get an attribution statement that can simply be copied and pasted in their "documents". Generally, the citation string has a generic format, e.g., DataCite style (Starr and Gastl, 2011), consisting of authors, year of publication, title, publisher, repository name, and DOI.
- *Export option* Allows users to directly export the citation to the dataset in a variety of generic formats. The most popular ones are RIS (compatible with reference management software such as EndNote, Reference Manager, ProCite, and RefWorks) and BibTex (compatible with software such as LaTeX and BibDesk), but others may be available, e.g., DataCite, Dublin Core, NLM, and MARCXML.
- *Embed option* Allows users to embed a link to the data in the HTML source code of their web pages.
- *Share option* Allows users to share a link to the dataset directly via email, e.g., Zenodo, or through a variety of social networks, such as Twitter, Facebook, Google Plus, Tumblr, and Mendeley.

It is worth noticing that all these approaches rely on the DOI that is assigned to every dataset at publication time.

	3TU.Dat.	CSIRO	Dryad	Figshare	Zenodo
Citation string	✓	✓	✓	✓	✓
Export option	✓		✓	✓	✓
Embed option	✓			✓	
Share option			✓	✓	✓

Table 9: Dataset citation practices supported by Scientific Data Repositories.

What is missing and trends. Data citation currently is subject of an intense research activity, e.g., CODATA-ICSTI Task Group on Data Citation Standards and Practices (2013). The research community has recently discussed and released eight principles on proper data citation (Data Citation Synthesis Group, 2014; Altman et al., 2015). These principles cover purpose, function and attributes of data citations. However they are "just" principles and require implementation guidelines. In fact, Starr et al. (2015) recently proposed a framework for operationalizing these principles. It can be affirmed that current data repositories meet the principles to a certain extent, but a lot still needs to be done to address detailed requirements that scientists have when publishing or using datasets. For example, repositories contribute to supply credit to data contributors, however they have not yet developed any micro attribution oriented facility aiming at highlighting who did what (Allen et al., 2014). Communities often have their own way to cite their data. It would be very useful that repositories offer facilities enabling specific communities to customise the way their own data should be cited preserving interoperability at the same time.

Very often datasets used in experiments and studies are obtained by issuing specific queries on dynamic sources like databases. Requiring scientists to upload static snapshots or textual descriptions of data subsets, as most of the current repository do, is not enough to enable precise identification of the specific portion and of the version of data, i.e., what is required for the reproducibility of processes, sharing and reuse. A recent proposal for dealing with these scenarios require that data sources and appropriate infrastructures (including repositories) are considerably modified (Rauber et al., 2015). In particular, it is proposed that data sources reinforce their query facilities thus to enable *data versioning*, i.e., ensure that earlier "forms" of the dataset can be retrieved, and *timestamping*, i.e., ensure that operations on datasets are marked with timestamp. Repositories dealing with data publishing can contribute to implement the proposed approach by supporting the phases of (*a*) persistently identifying the dataset, e.g., deal with the storage of the query univocally identifying the dataset portion and the relative metadata, assign the Persistent Identifier (PID) to this rich query, and (*b*) resolving the PIDs and actually retrieve the data portion, e.g., produce both a human-readable and machine-actionable landing page for accessing the dataset portion (via query execution) and the metadata. Thus repositories are expected to act like custodians of the specification characterising the dataset portion.

Another requirement related to data citation is to have machine readable citations, e.g., to enable "service providers" to build and operate "services" on it. Approaches exist and initiatives are ongoing to make them suitable for data. The Citation Typing Ontology (CiTO) (Peroni and Shotton, 2012) is an ontology to enable characterising citations. Some of its relationships can be used to deal with the data case, e.g., "cites as data source", "uses data from". Starr et al. (2015) have proposed an approach that by complementing the Journal Article Tag Suite (JATS), an XML schema for tagging journal articles for exchanging journal content, contributes to implement the data citation principles. They are proposing to have univocally identified and persistent landing pages that provide human-readable (namely HTML) and machine-readable (e.g., JSON) basic information on the dataset. Repositories are expected to start experiencing with these approaches.

In supporting data publishing practices, repositories should also take into account the emerging trend of documenting produced datasets through data papers (Candela et al., 2015). Data papers contain information about the datasets and how they have been produced, thus facilitating a better understanding of them. This largely contributes to simplify and enhance datasets re-use. In order to fully exploit their power, repositories should first of all maintain a link between the dataset and its data paper(s), if any, and encourage data users willing to cite the datasets to actually cite the data paper(s). Moreover, in order to have a more complete understanding, repositories should also maintain a link to all the paper(s) referring to the dataset since the relationship between datasets and papers are many and multiform (Borgman, 2015). Clearly, to realise the just suggested approach publishers must agree and develop both detailed guidelines and copy editing practices to guarantee that the data papers they publish contain an intelligible citation to the dataset(s)

each data paper is about. These citations should include a reference to the repository(ies) where the data is actually stored.

5 Conclusion and Prospect

This article has surveyed the practices and approaches for data publishing promoted by generalist scientific data repositories, i.e., repositories accepting the publication of any type of dataset. The survey has been conducted on repositories that are recommended by data journals. In so doing we are relying on a sort of certification by publishers and communities, i.e., the repositories are expected to have a sufficient level of quality in exposing data and making them accessible for validation and re-use. In applying these selection criteria we have necessarily excluded other available repositories and systems. In particular, we excluded Institutional Repositories that certainly play a key role is research data stewardship, e.g., Cragin et al. (2010) highlighted how certain datasets "can be managed well in an IR context when characteristics and practices are well understood". Among the excluded repositories there is B2SHARE (Berenji Ardestani et al., 2015). This is one of the facilities released by the EUDAT initiative and it is oriented to provide researchers with a user-friendly and reliable way to store and share research datasets. It is based on Invenio, the same system that is at the core of Zenodo. B2SHARE assigns a persistent identifier to each deposited dataset to ensure long-lasting access and reference. It is completely interoperable with other EUDAT developed services, B2FIND and B2SAFE, that provide discovery and availability over the time. Another example is represented by the Globus Data Publication facilities (Chard et al., 2015). This approach is building on DSpace (Smith et al., 2003) and other technologies to realize a data repository. These are two among the many recent initiatives that are dealing with data publishing, especially in the "eScience" domain.

The paper has presented the current practices of the selected repositories, and discussed their open issues and prospects with respect to eight key data publishing features, i.e., dataset formatting, documentation, licensing, publication costs, validation, availability, discovery and access, and citation. Two major observations have clearly emerged when analysing repositories' contribution to data publishing from these different perspectives: (*a*) the legacy of literature publishing and (*b*) the lack of a designated community.

The support that generalist data repositories currently offer mainly resembles established practices in literature publishing, very few facilities are rethought to deal with datasets peculiarities and their different usage. For example, relevant aspects regarding the scope of data, e.g., coverage, and contextual information, e.g., provenance and attribution, often do not receive the necessary attention. This discrepancy becomes even more evident when taking into account new scientific practices triggered by technological progresses in data management. Web-based collaborative environments, for example, are opening the way to the creation of data products through the collaborative effort of many scientists that produce and publish these products at different stages and, hence, in many versions. This need is not reflected in the features offered by the analysed repositories that pose little emphasis on the *versioning* issues (Kowalczyk and Shankar, 2011; Buneman et al., 2004). Their main focus is actually limited to the "final" datasets only. Similarly, systems supporting annotation and rating now enable associating a progressively growing amount of contextual information with the published dataset. Repositories are not yet designed to collect, make available and usefully exploit the so gathered information for providing, for example, better discovery and validation services.

Generalist repositories are potentially exposed to (*a*) a large variety of cases, e.g., an almost open ended set of dataset formats and typologies to deal with, (*b*) the largest and most multidisciplinary community of practice possible in terms of both data owners and data consumers, and (*c*) a lack of consolidated and shared practices. To deal with this variety, repositories currently tend to prefer consolidated approaches rather than investing effort in experimenting innovative solutions for supporting dataset publishing. It is expected that this is a transient steps and that the situation will rapidly evolve in the next years. The data publishing market is expected to grow very rapidly. All the analysed repositories, as many others, have made the choice of entering in the market at this initial stage, even if this means proposing a traditional product that only partially meets the complex needs of those that want to validate and re-use data. It is expected that in the years to come, when market positions will be more consolidated and competitive, repositories will invest more resources on innovative solutions providing feature-rich services for helping data publishing.

Thus we can conclude that current generalist repositories are viable services for data publishing yet need to evolve and reinforce their offering to better support it.

A key issue that generalist repositories will certainly have to address to overcome the limitations imposed by the extreme heterogeneity of the managed datasets is how to improve the specification that repositories offer of themselves, related to, e.g., the formats they natively support, the validation procedures offered

and the data licences managed. This information is today rarely accessible in detail even in human readable form. Data producers, editors imposing data deposition policies and consumers willing to select the most appropriate repository for their needs hardily find information sufficient to make their choice. More often other criteria, like trust, cost and usage by known stakeholders, guide their selection. When this descriptive specification will be available in automatically processable form the implementation of more tuned and adaptable repository services will became feasible, e.g., dataset pre-validation, for example, will be dynamically customizable on the set of formats that are accepted at a certain time. This specification will also facilitate third-party service providers that will be able, for example, to understand if the re-use of certain datasets is possible for them or how to better implement federated retrieval systems on top of the repository itself. Another major expectation concerns the interaction between repositories and their users, both data owners and data consumers, that will necessarily have to become more flexible, open, feature-rich and collaborative.

These needs have already emerged and there are early attempts to address them, e.g., by envisaging pre-publication repositories (Steinhart, 2007), by foreseeing the use of software management practices (Schopf, 2012; Gandrud, 2013), and by completely rethinking the publishing act (Assante et al., 2015). To some extent, the act of publishing a dataset in a repository should be just the beginning of large scale collaboration. Researchers other than the data owners can "contribute" to the dataset by (*a*) producing new versions of it, e.g., by enriching the content or correcting errors, (*b*) reporting their concrete experiences, both successful and unsuccessful, in reusing the dataset for a certain investigation, e.g., by publishing a research paper to be added to the dataset documentation, (*c*) linking it with other datasets that are complementary or incompatible with respect to a given goal. Generalist repositories are not alone in confronting with the data publishing goal. They are part of an ecosystem where diverse actors are called to spend effort to define and agree on shared policies, practices and roles guiding proper data publishing (Castelli et al., 2013; Borgman, 2015; Candela et al., 2015).

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The work reported has been partially supported by the *iMarine* project (FP7 of the European Commission, FP7-INFRASTRUCTURES-2011–2, Contract No. 283644) and the BlueBRIDGE project (European Union's Horizon 2020 research and innovation programme under the grant agreement No. 675680). The authors would like to thank the selected repositories team for their kind and informative replies to specific requests for details. The authors would like to thank M. B. Baldacci (ISTI-CNR) for her valuable support and the many helpful comments. The authors would like to thank both reviewers for their insightful comments on the paper.

Author contributions

Contributions to the paper are described using the taxonomy described in Allen et al. (2014). Study conception: LC DC. Methodology: LC DC. Investigation: Data collection: MA AT. Writing: initial draft: AT. Writing: LC. Writing: critical review: LC DC. Data curation: MA LC AT.

References

Abelson, H, Adida, B, Linksvayer, M and **Yergler, N** 2008 ccREL: The Creative Commons Rights Expression Language. *Tech. rep.*, Creative Commons.

Adie, E and **Roe, W** 2013 Altmetric: enriching scholarly content with article-level discussion and metrics. *Learned Publishing*, 26(1): 11–17. DOI: http://dx.doi.org/10.1087/20130103

Allen, L, Scott, J, Brand, A, Hlava, M and **Altman, M** 2014 (April) Publishing: Credit where credit is due. Nature, 312–313. Available at http://www.nature.com/news/publishing-credit-where-credit-is-due-1.15033. DOI: http://dx.doi.org/10.1038/508312a

Altman, M, Borgman, C L, Crosas, M and **Martone, M** 2015 An introduction to the joint principles for data citation. *Bulletin of the Association for Information Science and Technology*, 41(3). DOI: http://dx.doi.org/10.1002/bult.2015.1720410313

Andreoli-Versbach, P and **Mueller-Langer, F** 2014 Open access to data: An ideal professed but not practised. *Research Policy*, 43(9): 1621–1633. DOI: http://dx.doi.org/10.1016/j.respol.2014.04.008

Asher, A, Deards, K, Esteva, M, Halbert, M, Jahnke, L, Jordan, C, Keralis, S D C, Kulasekaran, S S, Moen, W E, Stark, S, Urban, T and **Walling, D** 2013 Research data management: Principles, practices, and prospects. *Tech. rep.*, Council on Library and Information Resources.

Assante, M, Candela, L, Castelli, D, Manghi, P and **Pagano, P** 2015 Science 2.0 repositories: Time for a change in scholarly communication. *D-Lib Magazine*, 21(1/2). DOI: http://dx.doi.org/10.1045/january2015-assante

Ball, A 2012 How to license research data. *Tech. rep.*, Digital Curation Centre.

Ball, A, Chen, S, Greenberg, J, Perez, C, Jeffery, K and **Koskela, R** 2014 Building a disciplinary metadata standards directory. *International Journal of Digital Curation*, 9(1): 142–151. DOI: http://dx.doi.org/10.2218/ijdc.v9i1.308

Beckerle, M J and **Hanson, S M** 2014 (September) Data Format Description Language (DFDL) v1.0 Specification. *Tech. Rep. GFD-P-R.207*, Open Grid Forum.

Berenji Ardestani, S, Håkansson, C J, Laure, E, Livenson, I, Straňák, P, Dima, E, Blommesteijn, D and **van de Sanden, M** 2015 B2SHARE: An open escience data sharing platform. In: 11th IEEE International Conference on eScience, Munich, Germany.

Berman, F 2008 Got data? a guide to data preservation in the information age. *Communications of the ACM*, 51(12): 50–56. DOI: http://dx.doi.org/10.1145/1409360.1409376

Bizer, C, Heath, T and **Berners-Lee, T** 2009 Linked data - the story so far. *International Journal on Semantic Web & Information Systems*, 5(3): 1–22. DOI: http://dx.doi.org/10.4018/jswis.2009081901

Bobadilla, J, Ortega, F, Hernando, A and **Gutiérrez, A** 2013 Recommender systems survey. *Knowledge-Based Systems*, 46: 109–132. DOI: http://dx.doi.org/10.1016/j.knosys.2013.03.012

Borgman, C 2011 The Conundrum of Sharing Research Data. *Journal of the Association for Information Science and Technology*, 63(6): 1059–1078. DOI: http://dx.doi.org/10.1002/asi.22634

Borgman, C L 2015 Big Data, Little Data, No Data: Scholarship in the Networked World. The MIT Press.

Bourne, P E 2010 What Do I Want from the Publisher of the Future? *PLoS Computational Biology*, 6(5): e1000787. DOI: http://dx.doi.org/10.1371/journal.pcbi.1000787

Bourne, P E, Clark, T, Dale, R, de Waard, A, Herman, I, Hovy, E H and **Shotton, D** 2012 Improving the future of research communication and e-scholarship. *Force11 white paper*, Force11.

Broeder, D and **Lannom, L** 2014 Data type registries: A research data alliance working group. *D-Lib Magazine*, 20(1/2). DOI: http://dx.doi.org/10.1045/january2014-broeder

Buneman, P, Khanna, S, Tajima, K and **Tan, W.-C** 2004 Archiving scientific data. *ACM Transactions on Database Systems*, 29(1): 2–42. DOI: http://dx.doi.org/10.1145/974750.974752

Burda, D and **Teuteberg, F** 2013 Sustaining accessibility of information through digital preservation: A literature review. *Journal of Information Science*, 39(4): 442–458. DOI: http://dx.doi.org/10.1177/0165551513480107

Campbell, J 2015 Access to scientific data in the 21st century: Rationale and illustrative usage rights review. *Data Science Journal*, 13: 203–230. DOI: http://dx.doi.org/10.2481/dsj.14-043

Candela, L, Castelli, D, Manghi, P and **Tani, A** 2015 Data journals: A survey. *Journal of the Association for Information Science and Technology*, 66(9): 1747–1762. DOI: http://dx.doi.org/10.1002/asi.23358

Carata, L, Akoush, S, Balakrishnan, N, Bytheway, T, Sohan, R, Seltzer, M and **Hopper, A** 2014 A primer on provenance. *Communications of the ACM*, 57(5): 52–60. DOI: http://dx.doi.org/10.1145/2596628

Castelli, D, Manghi, P and **Thanos, C** 2013 A vision towards scientific communication infrastructures. *International Journal on Digital Libraries*, 13(3–4): 155–169. DOI: http://dx.doi.org/10.1007/s00799-013-0106-7

CCSDS 2012 Reference model for an open archival information system. *Recommended practice CCSDS 650.0-M-2*, Consultative Committee for Space Data Systems.

Chard, K, Pruyne, J, Blaiszik, B, Ananthakrishnan, R, Tuecke, S and **Foster, I** 2015 Globus data publication as a service: Lowering barriers to reproducible science. In: 11th IEEE International Conference on eScience, Munich, Germany.

CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013 Out of cite, out of mind: The current state of practice, policy, and technology for the citation of data. *Data Science Journal*, 12: CIDCR1–CIDCR75.

Cornillon, P, Gallagher, J and **Sgouros, T** 2003 OPeNDAP: Accessing data in a distributed, heterogeneous environment. *Data Science Journal*, 2. DOI: http://dx.doi.org/10.2481/dsj.2.164

Costas, R, Meijer, I, Zahedi, Z and **Wouters, P** 2013 (April) The value of research - data metrics for datasets from a cultural and technical point of view. Knowledge exchange report, Knowledge Exchange. URL www.knowledge-exchange.info/datametrics

Costello, M J, Michener, W K, Gahegan, M, Zhang, Z-Q and **Bourne, P E** 2013 Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution*, 28(8): 454–461. Available at http://

www.sciencedirect.com/science/article/pii/S0169534713001092. DOI: http://dx.doi.org/10.1016/j.tree.2013.05.002

Cragin, M H, Palmer, C L, Carlson, J R and **Witt, M** 2010 Data sharing, small science and institutional repositories. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1926): 4023–4038. DOI: http://dx.doi.org/10.1098/rsta.2010.0165

Crosas, M 2011 The Dataverse Network: An Open-Source Application for Sharing, Discovering and Preserving Data. *D-Lib Magazine*, 17(1/2). DOI: http://dx.doi.org/10.1045/january2011-crosas

Daga, E, D'Aquin, M, Motta, E and **Gangemi, A** 2015 A bottom-up approach for licences classification and selection. In: Gandon, F., Gu´eret, C., Villata, S., Breslin, J., Faron-Zucker, C., Zimmermann, A. (Eds.), The Semantic Web: ESWC 2015 Satellite Events, Portorož, Slovenia, May 31 – June 4, 2015, Revised Selected Papers. Springer International Publishing, pp. 257–267. DOI: http://dx.doi.org/10.1007/978-3-319-25639-9_41

Data Citation Synthesis Group 2014 Joint Declaration of Data Citation Principles. Available at http://www.force11.org/datacitation, accessed August 2014.

Denenberg, R 2009 Search web services – the oasis sws technical committee work – the abstract protocol definition, opensearch binding, and sru/cql 2.0. *D-Lib Magazine*, 15(1/2). DOI: http://dx.doi.org/10.1045/january2009-denenberg

Devine, E 2014 (November) Making data beautiful: the importance of supplemental material. *Digital Science Blog*. Available at http://www.digital-science.com/blog/guest/making-data-beautiful-the-importance-of-supplemental-material/.

Dobratz, S and **Scholze, F** 2006 DINI institutional repository certification and beyond. *Library Hi Tech*, 24(4): 583–594. DOI: http://dx.doi.org/10.1108/07378830610715446

Douglass, K, Allard, S, Tenopir, C, Wu, L and **Frame, M** 2014 Managing scientific data as public assets: Data sharing practices and policies among full-time government employees. *Journal of the Association for Information Science and Technology*, 65(2): 251–262. DOI: http://dx.doi.org/10.1002/asi.22988

Eschenfelder, K R and **Johnson, A** 2014 Managing the data commons: Controlled sharing of scholarly data. *Journal of the Association for Information Science and Technology*, 65(9): 1757–1774. DOI: http://dx.doi.org/10.1002/asi.23086

figshare 2013 (January) figshare partners with open access mega journal publisher PLOS. *Figshare Blog*. Available at http://figshare.com/blog/figshare_partners_with_Open_Access_mega_journal_publisher_PLOS/68

Gandrud, C 2013 GitHub: A tool for social data set development and verification in the cloud. *The Political Methodologist*, 20(2): 7–16. DOI: http://dx.doi.org/10.2139/ssrn.2199367

Guha, R V, Brickley, D and **Macbeth, S** 2016 Schema.org: Evolution of structured data on the web. *Communications of the ACM*, 59(2), 44–51. DOI: http://dx.doi.org/10.1145/2844544

Guibault, L and **Wiebe, A** (Eds.) 2014 Safe to be open: Study on the protection of research data and recommendations for access and usage. Universitätsverlag Göttingen.

Heidorn, P B 2008 Shedding light on the dark data in the long tail of science. *Library Trends*, 57(2): 280–299. DOI: http://dx.doi.org/10.1353/lib.0.0036

Hotan, A W, van Straten, W and **Manchester, R N** 2014 PSRCHIVE and PSRFITS: An Open Approach to Radio Pulsar Data Storage and Analysis. *Publications of the Astronomical Society of Australia*, 21(3): 302–309. DOI: http://dx.doi.org/10.1071/AS04022

Iannella, R 2002 Open Digital Rights Language (ODRL) Version 1.1. W3c note, W3C. Available at http://www.w3.org/TR/odrl.

Kansa, E C, Kansa, S W and **Arbuckle, B** 2014 Publishing and pushing: Mixing models for communicating research data in archaeology. *International Journal of Digital Curation*, 9(1): 57–70. DOI: http://dx.doi.org/10.2218/ijdc.v9i1.301

Klump, J, Bertelmann, R, Brase, J, Diepenbroek, M, Grobe, H, Höck, H, Lautenschlager, M, Schindler, U, Sens, I and **Wächter, J** 2006 Data publication in the open access initiative. *Data Science Journal*, 5: 79–83. DOI: http://dx.doi.org/10.2481/dsj.5.79

Kowalczyk, S and **Shankar, K** 2011 Data sharing in the sciences. *Annual Review of Information Science and Technology*, 45(1): 247–294. DOI: http://dx.doi.org/10.1002/aris.2011.1440450113

Kratz, J 2014 (May) Fifteen ideas about data validation (and peer review). Data Pub Blog. Available at http://datapub.cdlib.org/2014/05/08/fifteen-ideas-about-data-validation-and-peer-review/.

Kratz, J and **Strasser, C** 2014 Data publication consensus and controversies. *F1000Research*, 3(94). [v3; ref status: indexed, http://f1000r.es/4ja]. DOI: http://dx.doi.org/10.12688/f1000research.3979.3

Lagoze, C and **Van de Sompel, H** 2001 The open archives initiative: building a low-barrier interoperability framework. In: Proceedings of the first ACM/IEEE-CS Joint Conference on Digital Libraries. ACM Press, pp. 54–62. DOI: http://dx.doi.org/10.1145/379437.379449

Lawrence, B, Jones, C, Matthews, B, Pepler, S and **Callaghan, S** 2011 Citation and peer review of data: Moving towards formal data publication. *International Journal of Digital Curation*, 6(2): 4–37. DOI: http://dx.doi.org/10.2218/ijdc.v6i2.205

Manghi, P, Artini, M, Atzori, C, Bardi, A, Mannocci, A, La Bruzzo, S, Candela, L, Castelli, D and **Pagano, P** 2014 The D-NET software toolkit – a framework for the realization, maintenance, and operation of aggregative infrastructures. *Program: electronic library and information systems*, 48(4): 322–354. DOI: http://dx.doi.org/10.1108/PROG-08-2013-0045

Manghi, P, Bolikowski, L, Manola, N, Schirrwagen, J and **Smith, T** 2012 OpenAIREplus: the european scholarly communication data infrastructure. *D-Lib Magazine*, 18(9/10). DOI: http://dx.doi.org/10.1045/september2012-manghi

Manghi, P and **Mannocci, A** 2013 Data Searchery: Preliminary Analysis of Data Sources Interlinking. In: Research and Advanced Technology for Digital Libraries. Vol. 8092 of Lecture Notes in Computer Science. pp. 458–461.

Mannheimer, S, Yoon, A, Greenberg, J, Feinstein, E and **Scherle, R** 2014 A balancing act: The ideal and the realistic in developing Dryad's preservation policy. *First Monday*, 19(8). DOI: http://dx.doi.org/10.5210/fm.v19i8.5415

Marcial, L H and **Hemminger, B M** 2010 Scientific data repositories on the web: An initial survey. *Journal of the American Society for Information Science and Technology*, 61: 2029–2048. DOI: http://dx.doi.org/10.1002/asi.21339

Mayernik, M S, Callaghan, S, Leigh, R, Tedds, J and **Worley, S** 2014 Peer review of datasets: When, why, and how. *Bulletin of the American Meteorological Society e-View*.

McGath, G 2013 The format registry problem. *Code4Lib*, 19.

Nicol, A, Caruso, J and **Archambault, É** 2013 (August) Open data access policies and strategies in the european research area and beyond. *Tech. rep.*, Science- Metrix Inc.

OECD 2007 OECD Principles and Guidelines for Access to Research Data from Public Funding. OECD Publications.

Palmer, C L, Cragin, M H, Heidorn, P B and **Smith, L C** 2007 Data curation for the long tail of science: The case of environmental sciences. In: Third International Digital Curation Conference, Washington, DC.

Palmer, C L, Weber, N M and **Cragin, M H** 2011 The analytic potential of scientific data: Understanding reuse value. *Proceedings of the American Society for Information Science and Technology*, 48(1): 1–10. DOI: http://dx.doi.org/10.1002/meet.2011.14504801174

Pampel, H and **Dallmeier-Tiessen, S** 2014 Open research data: From vision to practice. In: Bartling, S., Friesike, S. (Eds.), *Opening Science*. Springer International Publishing, pp. 213–224. DOI: http://dx.doi.org/10.1007/978-3-319-00026-8_14

Parsons, M and **Fox, P** 2013 Is data publication the right metaphor? *Data Science Journal*, 12: WDS31–WDS46. DOI: http://dx.doi.org/10.2481/dsj.WDS-042

Parsons, M A and **Duerr, R** 2005 Designating user communities for scientific data: challenges and solutions. *Data Science Journal*, 4: 31–38. DOI: http://dx.doi.org/10.2481/dsj.4.31

Parsons, M A, Gødoy, Ø, LeDrew, E, de Bruin, T F, Danis, B, Tomlinson, S and **Carlson, D** 2011 A conceptual framework for managing very diverse data for complex, interdisciplinary science. *Journal of Information Science*, 37(6): 555–569. DOI: http://dx.doi.org/10.1177/0165551511412705

Peroni, S and **Shotton, D** 2012 FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. *Web Semantics: Science, Services and Agents on the World Wide Web*, 17: 33–43. DOI: http://dx.doi.org/10.1016/j.websem.2012.08.001

Peters, D 2015 (June) Wiley partnership with figshare enables data sharing. Press release. Available at http://eu.wiley.com/WileyCDA/PressRelease/pressReleaseId-119082.html.

Piwowar, H A 2013 Altmetrics: Value all research products. *Nature*, 493(159).

Rauber, A, Asmi, A, van Uytvanck, D and **Proll, S** 2015 (September) Data citation of evolving data. RDA working group on Data Citation: Making Dynamic Data citeable Recommendations https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_150924.pdf.

Raymond, E S 2008 The Art of UNIX Programming. Addison-Wesley.

Renear, A H, Sacchi, S and **Wickett, K M** 2010 Definitions of dataset in the scientific and technical literature. *Proceedings of the American Society for Information Science and Technology*, 47(1): 1–4. DOI: http://dx.doi.org/10.1002/meet.14504701240

Roche, D G, Jennions, M D and **Binning, S A** 2013 Data deposition: Fees could damage public data archives. *Nature*, 502(7470): 171. DOI: http://dx.doi.org/10.1038/502171a

Rombouts, J and **Princic, A** 2010 Building a 'data repository' for heterogenous technical research communities through collaborations. In: International Association of Scientific and Technological University Libraries, 31st Annual Conference.

Rousidis, D, Garoufallou, E, Balatsoukas, P and **Sicilia, M-A** 2014 Data Quality Issues and Content Analysis for Research Data Repositories: The Case of Dryad. In: ELPUB2014. Let's put data to use: digital scholarship for the next generation, 18th International Conference on Electronic Publishing 1920 June 2014, Thessaloniki, Greece. IOS Press, pp. 45–98.

Schopf, J M 2012 Treating Data Like Software: A Case for Production Quality Data. In: Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries. ACM, pp. 153–156. DOI: http://dx.doi.org/10.1145/2232817.2232846

Simmhan, Y L, Plale, B and **Gannon, D** 2005 (Sep.) A survey of data provenance in e-science. *SIGMOD Rec.* 34(3): 31–36. Available at http://doi.acm.org/10.1145/1084805.1084812. DOI: http://dx.doi.org/10.1145/1084805.1084812

Smith, M, Barton, M, Bass, M, Branschofsky, M, McClellan, G, Stuve, D, Tansley, R and **Walker, J** 2003 DSpace – An Open Source Dynamic Digital Repository. *D-Lib Magazine*, 9(1). Available at http://www.dlib.org/dlib/january03/smith/01smith.html. DOI: http://dx.doi.org/10.1045/january2003-smith

Starr, J, Castro, E, Crosas, M, Dumontier, M, Downs, R R, Duerr, R, Haak, L L, Haendel, M, Herman, I, Hodson, S, Hourclé, J, Kratz, J E, Lin, J, Nielsen, L H, Nurnberger, A, Proell, S, Rauber, A, Sacchi, S, Smith, A, Taylor, M and **Clark, T** 2015 Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science*, 1(e1).

Starr, J and **Gastl, A** 2011 isCitedBy: A metadata scheme for DataCite. *D-Lib Magazine*, 17: n.a. DOI: http://dx.doi.org/10.1045/january2011-starr

Steinhart, G 2007 Datastar: An institutional approach to research data curation. *IASSIST Quarterly*, 31(3/4): 34–39.

Tenopir, C, Allard, S, Douglass, K, Aydinoglu, A U, Wu, L, Read, E, Manoff, M and **Frame, M** 2011 Data sharing by scientists: Practices and perceptions. *PLoS ONE*, 6(6): e21101. DOI: http://dx.doi.org/10.1371/journal.pone.0021101

Thanos, C 2014 Scientific data reusability: Conceptual foundations, impediments and enabling technologies. *Tech. rep.*, Istituto di Scienza e Tecnologie dell'Informazione "A. Faedo", CNR.

Wang, X, De Martini, T, Wragg, B, Paramasivam, M and **Barlas, C** 2005 The MPEG-21 rights expression language and rights data dictionary. *IEEE Transactions on Multimedia*, 7(3): 408–417. DOI: http://dx.doi.org/10.1109/TMM.2005.846788

Willis, C, Greenberg, J and **White, H** 2012 Analysis and synthesis of metadata goals for scientific data. *Journal of the American Society for Information Science and Technology*, 63(8): 1505–1520. DOI: http://dx.doi.org/10.1002/asi.22683

How to cite this article: Assante, M, Candela, L, Castelli, D and Tani, A 2016 Are Scientific Data Repositories Coping with Research Data Publishing? *Data Science Journal,* 15: 6, pp. 1–24, DOI: http://dx.doi.org/10.5334/dsj-2016-006

Submitted: 25 November 2015 **Accepted:** 05 April 2016 **Published:** 26 April 2016

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

OPEN ACCESS