Do researchers use Open Research Data? Exploring the relationships between usage trends and metadata quality across scientific disciplines from the Figshare case

Auth1[1] Auth2 [2]

Abstract

Open Research Data (ORD) have been considered a driver of scientific transparency. However, data friction, as the phenomenon of data underutilization for several causes, has also been pointed out. A factor often called into question for ORD low usage is the quality of the ORD and associated metadata. This work aims to illustrate the use of ORD, published by the Figshare scientific repository, concerning their scientific discipline, their type and compared to the quality of their metadata. Considering all the Figshare resources, and carrying out a programmatic quality assessment of their metadata, our analysis highlighted two aspects. Firstly, irrespective of the scientific domain considered, most ORD are under-used, but with exceptional cases which concentrate most researchers' attention. Secondly, there was no evidence that the use of ORD is associated with good metadata publishing practices. These two findings opened to a reflection about the potential causes of such data friction.

Keywords

Open Research Data usage, metadata quality, research repositories

Introduction

The Open Research Data (ORD) movement can be collocated within the broader context of Open Science, which advocates for public and accessible science[1,2]. The movement has evolved to embrace new researchers' practices and identities, beyond the idea of a digital science and towards open and social activities which entail international collaboration in increasingly complex data infrastructures[3]. In fact, it has been pointed out that the effectiveness, productivity and reproducibility of scientific findings is deeply linked to sharing and reusing open research data[4]. According to Molloy[5] "The more data is made openly available in a useful manner, the greater the level of transparency and reproducibility" (Molloy,2011, p.1). In this same vein, for[6] transparency in research can be placed into a three dimensional model of Open science (together with participation and knowledge access). It must be considered "as an outcome from a combination of different behaviours and practices associated with reproducibility" (Lyon, 2016, p.161). However, the attitude toward making research transparent varies across disciplines, with domains such as astronomy and genomics more advanced in following open data practices than, for instance, humanities and social sciences[6].

Also Borgman[7] has pointed out that researchers' practices differ from discipline to discipline over the basis of different data approaches and professional cultures. Even if ORD are considered a driving force for transparency and research effectiveness[6,8,9], these have limited value if they are not utilized[10]. In this regard, standard parameters to produce ORD, ensuring their quality has been considered of paramount importance to support sharing and reusing activities[11]. The quality of the datasets (i.e. *open data resources*) and associated metadata is often suggested as a precondition for the success of data openness programs not only in research but also in all public data generated by eGovernment approaches.

[1] Affil1

[2] Affil2

Quality is indeed deemed as one of the possible causes that may hinder or prevent the users to effectively exploit Open Government Data (OGD) resources[12–14]. Specifically, the quality of metadata is seen as crucial as they are the bricks of open data catalogues, enabling users to search and consult the descriptions of the datasets, potentially improving the speed and easy of open data use[15–17]. In the case of ORD, good metadata are also essential to allow exploitation of datasets, by making them Findable, Accessible, Interoperable and Reusable either to human and by machines according to the FAIR principles[11]. However as observed by Sadiq and Indulska in[18], "the relationship between data quality (...) and the effective use of data remains unexplored in academic literature" (p.153). The same authors also point out that "there is a critical need (...) for empirical testing to identify the contexts and factors that affect the effectiveness of open data use" (p.153).

The literature on researchers professional practices around Social Media, as well as their digital skills (in terms of Digital Scholarship) has been investigated in depth in the last fifteen years[5,19]. Moreover, the progressive adoption of ORD and the pitfalls encountered while so doing, also raised attention[20]. However there is still shorthand of research on the social activity connected to minimal patterns of ORD reusability, for this phenomena would be more recent and more difficult to analyse. Bridging the research on the quality of ORD metadata and social media usage as part of digital scholarship, it can be assumed that poor usability patterns could be linked to the ORD quality and the researchers' skills to produce it. As[21] have pointed out, data collaboration is heavily dependent on contextual conditions of production of scientific work. The issues connected to such conditions lead to what the authors have called 'data friction', or the attrition between several data cultures determining not only data publication but the data description or metadata. This last indeed makes the contextual conditions clearer and transparent, encompassing less data friction. Therefore, data sharing not only depends on the quality of the data itself, but also the approach to metadata amongst the several scientific disciplines. Apart of the mentioned literature, prior research work on both authors supported such assumption. Previous work by[22] provided a first evaluation of the use of governmental datasets, based on a sample of national OGD portals. Similarly, in the case of[23], a small sample from the field of Educational Technologies on most important

Open Data Portals and repositories, yielded information about usage like downloading, viewing and citing ORD in connection with ORD quality. From such analysis we drawn two conclusions: i) most of the portals and repositories examined lack crucial if any information on the use of the published ORD; ii) for the few portals and repositories that publish usage data (mainly number of views and downloads) it can be deduced that most of the ORD are largely unused.

Guided by these observations, in this paper we take a step forward in characterising the users' engagement with Open Research portals, which not only collect ORD as complex resources always including machine readable datasets and metadata; but a number of other research resources like journals, figures, subsidiary tables and so one. After collecting programmatically via API, the metadata associated with about 7,000,000 open research resources published by Figshare, we analyzed the possible relationship between the quality of the metadata and the use of ORD.

The first contribution of this work is an updated photograph on the use of Figshare datasets across scientific fields. The main result confirms what we previously noted[22,23] , that is, in spite of a growing trend of ORD publication, most of them are scarcely used, albeit with some differences between the disciplinary fields. All in all, it was observed a concentration of usage over few datasets, in an unbalanced situation where most ORD are never consulted once published. The second contribution is the assessment of figshare ORD metadata quality though the adoption of a specific tool[*]. Overall metadata quality is seemingly appropriate. The third contribution is the analysis of the relationship between ORD usage and the quality of their metadata. Contrary to what is expected, our findings highlighted no preference for the resources of highest quality.

Background

Open Research Data

Open Data acquired a crucial importance for Open Science as object of socialization and exchange. The concept realizes concretely the conceptual ideals of openness in science[5]. Several international organizations are progressively covering data sharing in their funded projects, as in the international and European policy context the debate

---

[*]https://github.com/sebneu/portalwatch

achieved increasing relevance[20]. We could highlight several cases such as: the OpenAire portal as base for the visibility of Open Data coming from the European research framework Horizon2020[24]; the case of Wellcome Trust's policy states[25]; the Netherlands Organization for Scientific Research (NWO)[26]; the CERN's policies[27]; or the Bill Melinda Gates Foundation as private case[28]. For the European Commission, the centrality of Open Data has become a reality throughout the Mallorca Declaration of 2016[29]. Moreover, in the case of emerging economies, Open Access is also achieving more and more relevance as public policy given the expanded possibilities of researchers of the region to participate in the global science[30].

Overall, the policy documents claim that publicly funded research should be publicly available and accessible. However, it is also highlighted the potential of Open Data to endow the researchers to replicate the scientific work and to be reused under an economy of research resources[31]. Most importantly, Open Data can be mined by the industry ending up in innovations in faster cycles of RD[32]. The pressure to open data in science has become more and more relevant in parallel with the call for open data in all public activities and particularly in the context of eGovernment, being these activities maintained through public funding[33]. There are cross-fertilizations amongst these two movements, indeed. As it was pointed out in the OECD working paper from 2018,[34] there are two main reasons to consider the interactions between Open Government Data (OGD) and ORD. Firstly, the OGD can be of importance for the advancement of social sciences, humanities, health care and environmental sciences, which can directly adopt such data. Secondly, being ORD a specific category of public open data, the legal and ethical frameworks applied to OGD can often be directly applied to research data, particularly if this is produced with public funds. In such context, the research over the production, use and quality of OGD can cross-fertilize, with cautious considerations, to the ORD. As Dai, Shin and Smith (2018) point out "The Work on indicators and monitoring and impact assessment of OGD can also provide insights for equivalent issues relating to open science data" and "A culture of OGD needs to be nurtured in both the government and across the entire OGD ecosystem of users, including researchers" (Dai,Shin and Smith, 2018, p.14).

Along the increasing importance given to OGD and to ORD in policies, in the last field the investigations advanced in understanding the researchers engagement with Open Data publication in data repositories and portals[32]. However, the most enthusiastic discourses on the availability of data and the feasibility of appropriation by both the civil society (OGD) and researchers (ORD) immediately encounter a skills gap when dealing with activities such covering advanced data practices, crowd science, analysis of research quality and second hand data usage for industry or research purposes. In fact, some have compared the problem of appropriation of open data to the phenomenon of digital divide[35]. The research on overall data usage focused hence on mapping technology acceptance, patterns of usage and appropriation[36]. For the specific case of ORD, the lack of appropriate metadata explaining complex structures of data, as well as the lack of access or interoperability, or even machine readability of data, implies technical issues that hinder usage. The idea of developing the FAIR ("Findability, Accessibility, Interoperability, Reusability") data principles is an expression of the endeavor of introducing clear parameters for open data associated not only to human but also to machine tasks throughout algorithms and workflows[11].

Nonetheless, another crucial factor for the ORD usage is researchers data literacy, which seems to be still a concern. As a matter of fact, early in 2013[37] considered the need of generating a framework to address research data literacy.[38] presented a survey of data practices given to the Purdue College of Agriculture. The results showed rather basic data usages with no consideration of technical support from the Libraries.[39] presented the results of a quantitative study in Poland conducted within the framework of the international research project named ReDaM coordinated by the Information Literacy Association (InLitAs) in 2017. The results revealed that a significant number of respondents knew some basic concepts around Research Data Management (RDM), but they had not used institutional solutions elaborated in their parent institutions.[40] studied the information behaviour of Slovenian researchers in all research disciplines in relation to selected demographic variables through an online survey delivered to a random sample of the central registry of all active researchers in Slovenia. Age and discipline, and in a few cases also sex where noticeable factors influencing the researchers' information behaviour including

data management, curation and publishing within digital environments. Considering this situation, other studies focused the development of data literacy. For example,[41] early noticed that the researchers need to integrate the disposition, management and curation of data along research activities. Through a number of interviews and advanced students performance in activities of geoinformatics, the authors dealt with what they called the data information literacy program (DIL) preparing to achieve such needed skills.[42] considered the difficulties of finding good training resources for researchers and developed an introductory two-day intensive workshop on 'Data Carpentry', designed to teach basic concepts, skills, and tools for working more effectively and reproducibly with data.

This situation is highlighting the fact that the potential embedded in open data in science could not be directly transformed into effective practices towards Open Science if the researchers' skills gap is not covered. As a matter of fact, the practices around Open Data are unevenly distributed across scientific fields, and in most areas the concepts, tools and techniques to share data are little known[7,31]. To this regard, McKiernan et al. studying the literature until 2016 attempted to show the several benefits of sharing data in Applied Sciences, Life Sciences, Math, Physical Science and Social Sciences, where the advantages relate the visibility of research in terms of relative citations rates. In "The State of Open Data Report" by Springer[43] over 2,300 respondents reported also increasing numbers for the the publication of open data, but monitored only the intention (not the actual behavior) of specific forms of reuse and sharing. The same report in 2019[44] over 8,500 responses, reported the steady increase in awareness on ORD, their quality principles; moreover, it purports the willingness in several research fields and publishers to achieve open data as compulsory and rewarding practice for career advancement. However, the report was based on self-reported measures. When taking a look at the actual situation of Open Data trends at the Open Science Monitor of the European Commission[45], a sample pulled from the meta-portal R3Data* of OD repositories, in 2019, shows huge disparities between Life Sciences (1,295 OD repositories found), Natural Sciences (1,197), Humanities and Social Sciences (797) and Engineering (405). Furthermore, the European Data Portal concerning Open Government

---

*https://www.re3data.org

Data shows that there are also differences across disciplines in exploiting the potential of Open Data for research and innovation [46]. As the authors report "The European Data Portal (EDP) data category offering most mapped data sets is the justice, legal system public safety category (27.8%), followed by environment (23.6%), regions cities (12.0%), science technology (11.9%) and population society (5.5%). The category government public sector provides only 3.6 of the total mapped data sets while the economy finance category only provides 4.4" (p.26-27). While some of the problems could be linked to the quality of data published by the public administrations, as the authors claims, there is room to consider data literacy issues and substantial differences in research data practices across research domains as Borgman pointed early in 2015.

All in all, there is increasing attention coming from policy making as well as evidence on the advantages of sharing data. Notwithstanding, as we pointed out in this brief review of the literature these factors do not align with current practices. Furthering the research on mapping open research usage seems a yet important endeavor, going beyond self-reported measures which mostly purport intention but not actual behavior.

Open research data platforms

Platforms that support open research data have many points in common with those used by Open Government data portals to issue public domain data. Open Data platforms such as CKAN[*] and the commercial Socrata[†], enable data managers to release their datasets, assigning specific metadata with which to organize them into categories. These metadata allow users to recover the datasets of their interest through more or less advanced search features. They also provide APIs with which programmatically query the portals to download both metadata and datasets [47].

Open Research portals and repositories are frequently based on platforms which differ instead from government ones both for the type of published data and the actors involved in the publication and reuse of data. These platforms essentially publish research data and other research resources. These products need proper domain-specific metadata to be correctly interpreted and reused by other researchers from the same scientific fields. Researchers and research organizations are the

---

[*]https://ckan.org/features
[†]https://www.tylertech.com/products/socrata

main suppliers and beneficiaries of this data. It is thanks to their direct knowledge of the process and scientific context that led to the creation of the data, that its metadata can be fully described and made available to others[48]. As for the services and features, most platforms offer: storage capacity on the cloud per researcher; fields to define relevant metadata (scientific field and subfield, type of resource uploaded, type of licence, type of file, abstract, keywords, tags, etc.); post-usage services like sharing to social media, DOI, links of access, metrics of usage collection (views, downloads, citations, altmetrics). It's important to mention that the completion of the metadata fields are frequently self-defined by the author/researcher, encompassing eventual biases or cross-labeling when the resource comes from interdisciplinary research.

Metadata quality for Open Data

In Figshare's annual report, 'The State of Open Data 2018', Ross Wilkinson claims that "Increasingly funders of research are requiring verifiable quality", adding that "The quality of the data has to be able to be assessed"[13]. That data quality affects information retrieval, knowledge discovery and data reuse is known from early database management systems[49].

According to Edwards et al.[21], Science friction encompasses low reuse of datasets even by those who have not participated in their creation. This problem could be the resultant of weak quality metadata. In fact, the release of proper metadata depends on a series of factors that may limit its function as datasets' reuse enabler. Among these, the lack of common standard metadata has traditionally been included. However, this issue was subsequently remedied with the creation of specific metadata schemes, such as in the case of disciplines such as climatology (NetCDF CF) and environmental sciences (EML). Originally designed for documenting government data, W3C DCAT is a vocabulary for publishing data catalogs on the Web. Another factor that can cause the publication of poor quality metadata comes from those data repositories that allow data providers to set up their self-cured metadata, often combined with the difficulties of researchers to re-construing their study. As observed by Bates[17], and as we also reported for the Open Research Data section, the lack of skills or the unavailability of adequate supporting tools to prepare datasets' metadata may hamper the circulation of datasets. Scientific repositories such as

Fighsare, Dataverse, and Zenodo, have partially remedied this problem, imposing a pre-established structure and the presence of keywords that guide the scientist in documenting the resources deposited. However, the researcher remains responsible for completing the correct compilation of all the required metadata fields.

As data quality is acknowledge as a "multifaceted concept"[50] involving different dimensions (e.g. correctness, completeness, relevancy, availability, consistency), several methodological frameworks and technological proposals have been defined to assess various quality dimensions[51–56]. Some of these works, more recently, focused on evaluating OD portals taking into exam the quality of the data as well as of the associated metadata at different administrative level.

Reiche and Höfig[15] have defined and implemented five quality metrics for comparing the quality performance of three CKAN-based portals Germany, United Kingdom and UE. The authors recommended the creation of evaluation platforms to monitor OGD metadata. Neumaier, Umbrich and Pollere[16] have realized a metadata quality framework able to assess OGD portals based on various platforms. The framework maps different metadata profiles to the W3C DCAT metadata vocabulary and implements fifteen metrics on this uniform metadata structure. This allows the computation of OGD quality at the dataset level. A Web platform 'Open Data Portal Watch'[*] periodically assess the quality of 278 open data catalogues worldwide. We leveraged on, and tailored, the platform implementation to operate the metadata quality assessment of Figshare. Oliveira et al.[57] after having automatically and manually assessed metadata quality of 13 Brazilian OGD portals at the federal, regional and municipal administrative level, found that most of the evaluated datasets lacked metadata or presented a naive version of them. Vetró et al.[58] built up an assessment framework consisting of a set of 14 quality metrics to be applied both automatically and manually at the portal, dataset, and cell level. Authors applied their framework to two OGD portals samples which follow respectively a decentralized (i.e. non-common data structure) and centralized (i.e. with standardized data structures) disclosure strategy. Their findings, although on a rather small sample, show that centralized data disclosure provides better quality

---

[*]https://data.wu.ac.at/portalwatch

profiles. Máchová and Lnenicka[10] carried out a questionnaire-based quality assessment of 67 national portals, evaluating 28 quality criteria. From results analysis, authors suggest CDOs (Chief Data Officer) to introduce quality assessment practices for their portals, thus improving datasets delivery and increasing their findability and reusability. A 'User Interaction Framework', including 30 quality criteria, is presented by Zhu and Freeman[59] and operationalized through a coding book, for the evaluation of 34 US municipal Open Data portals. From the analysis of the results, the authors posit that more research has to be carried out to understand portals users' dynamics and intents. Based on the FAIR principles[11], Wilkinson et. al. designed an evaluation service framework, that implements 22 Maturity Indicators Tests, to measure the FAIRness of a Web resource[60]. These tests are grouped by the four principles in eighth (Findable), five (Accessible), seven (Interoperable) and two (Reusable) respectively. The FAIR Evaluation Services tool[*] allows users to select the whole 22 FAIR metrics, or one of the four subgroups, for assessing the FAIRness of a given (Web) resource by supplying its Globally Unique Identifier (GUID). At the end of the evaluation a summary of the assessment lists the successes and failures of the resource with respect the selected metrics. Users may interact with the evaluator either directly by Web interface or via APIs.

Material and Methods

This study aims at analysing basic parameters of usage of ORD according to their relationship with the quality of metadata and the scientific field, in order to further the evidence on the advancement on Open Data as essential piece of the Open Science. In order to achieve this aim, we designed a exploratory study focused on the analysis of the relationship between ORD publication, usage and the quality of the metadata.

Accordingly the research questions addressing this study are: 1- Which are the forms of publication and usage of ORD, considering the scientific fields of Figshare?; 2 - Which is the metadata quality of ORD in Figshare?; 3 - Does the metadata quality influence the forms of ORD usage in Figshare? In order to cover the above introduced research questions, in

---

[*]https://fairsharing.github.io/FAIR-Evaluator-FrontEnd

the following sections we introduce the instruments and metrics adopted to gather data, as well as the data analysis methods.

Usage Metrics To get basic insights on the data demand by users we analysed two indicators: the number of online views and the number of downloads associated to every portal datasets[61,62]. Several approaches to usage could be considered, but objective studies considering huge samples to characterize the situation across disciplines, would require a handful of parameters that can characterize some basic research data usage trends. In this regard, Scientometrics supports methodological approaches considering citations and altmetrics for several types of studies[63]. In the case of Social Media research, views, clicks and downloads also address the analysis of users' interactions and usage of digital content[64]. We mean by *Views* the number of times the page of a dataset was loaded in users' browsers and by *Downloads* the number of times a user has clicked (on URL or on a 'Download' button) to retrieve a file for a particular resource. These values can be returned by portal APIs and can be provided, along other metadata, as usage statistics in many open access repository platforms[65]. In some way, this information basically accounts for the activities of *direct users*, i.e. those who access the datasets directly[66]. There are other measures in the case of ORD like citations and altmetrics that can be considered more mature. In the case of OGD, the focus of attention has been put earlier than in the case of ORG on the assessment of datasets' impact on what could be considered the *indirect* users, who reuse OGD resources in further developments (apps, services, etc; e.g. the French* and Portuguese† national OGD portals introducing the number of applications that exploit each dataset, as associated metadata). These could be also the case of ORD, even if the field is in its infancy. In fact, most ORD platforms do not allow crawling advanced measures (citations, altmetrics) without accessing each ORD record (see next section). Moreover, intermediate processing or indirect use is not reported. For this reason we have limited ourselves to recovering direct access measures. Moreover, to characterise further the ORD usage, we considered other two parameters: scientific fields and ORD quality. As for the first one, we took the metadata 'Featured Categories' which

---

*https://www.data.gouv.fr/en/reuses
†https://dados.gov.pt/en/datasets/?sort=-reuses

are the scientific fields as self-reported (by the researcher). In spite that the self-reporting measures bias cannot be neglected, the researchers frequent association of own work to disciplinary areas lead them to be rather aware of such categories. We did not considered subcategories, which could imply further levels of subjective choice. Our findings relate Figshare scientific fields as categories and we never associate results to national or international nomenclatures. In the case of the second element (quality) we defined it prior in the background and further procedures to assign quality values have been described in the section "Metadata quality assessment".

Open research platforms and usage APIs To assess interest in research data, we initially explored the affordances and metrics of three widely used platforms: Zenodo*, Dataverse†, and Figshare‡, first of all, evaluating the presence of the two use indicators mentioned above. Dataverse reports download information only after the dataset has been selected by the user, from the list of datasets available in the catalog. There is, therefore, no way for the user to select the datasets of her interest based on an immediately visible popularity criterion. In addition to the number of downloads, Zenodo also displays the number of views, however only after the dataset has been selected by the user from the available ones. Unlike Dataverse, it is possible to sort the results of a search by popularity (Most Viewed). Figshare also adds the number of citations to views and downloads, however as well as the other two platforms, this information is available only after selecting the resource. Moreover, it is not included consistently as metadata.

After having verified the presence of some usage information, we ascertained how this data is made available via APIs. Another relevant information is the presence of categories that can be connected to research fields, even if self-selected by the same researchers. In any case,programmatic access to this essential information was included to conduct a systematic analysis of the use of ORD. Although the Dataverse documentation reports the possibility of accessing usage statistics at the

---

*https://zenodo.org/
†https://dataverse.harvard.edu/
‡https://figshare.com/

dataset level, via APIs*, repeated attempts have proved unsuccessful. After having contacted the support center (May 05th, 2019),the authors were informed that no available API endpoints turned on in Harvard Dataverse or the Demo Dataverse at the moment. Therefore, APIs were not expected to work. Having experienced similar issues with Zenodo APIs, on April 27th, 2019, the authors were informed that usage data was not available through the API at the moment, even if planed in the future.

Figshare usage information gathering Due to the limitations of programmatically gathering usage data occurring to Zenodo and Dataverse, our analysis of the use of open science repositories has focused on Figshare. Currently, Figshare manages more than 20 types of open research resources, e.g.: figures, journal contributions, technical reports, datasets, etc. Figshare resources are classified according to 21 scientific disciplines, ('featured categories'), which in turn are organized into a variable number of sub-categories (see Table 1). These scientific fields can be self-selected by the authors/researchers. Access via API to the metadata of Figshare ORDs takes place differently from that normally provided in platforms such as CKAN, Socrata. The latter provides APIs which directly gather the overall number of datasets published in a portal and their identifiers. In this way, the operations of metadata recovery are simplified, because it is always known at any moment what is the precise extent of a repository. To make up for this shortcoming, it is necessary to first identify the sub-categories of each discipline, then query for the associated resources identifiers, iteratively. This functionality is provided by a specific API† that implement the standard metadata extraction protocol OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting). Through the resource identifier, it is, therefore, possible to obtain individual metadata through the resource discovery API of Figshare. Without going into too much technical detail, let's just say that the current version of the API, v2, recommended in the documentation‡, does not include the usage data in the returned metadata, which are instead returned from the previous version, v1. However, the latter has less rich metadata. For example, it does not include license

---

*http://guides.dataverse.org/en/latest/api/native-api.html#dataset-metrics-api

†https://docs.Figshare.com/old_docs/OAI-PMH/, last accessed, 03/03/2020

‡https://docs.Figshare.com, last accessed, 03/03/2020

information. To remedy this situation, we have for each resource*, first used v1 to retrieve the usage data: views, and downloads (the citations are not present). A second call to v2 allowed retrieving all the other information needed to calculate the quality of the resource's metadata. Considering that, when we collected the data (2019/10/09-2019/11/23), the Figshare repository contained about seven million resources, the process required some weeks to complete.

Metadata quality assessment As for the quality assessment tool, we relied on the 'Open Data Portal Watch' platform code†, which implements the metadata quality methodology and metrics defined in [16]. The tool has meant to profile metadata quality of OD portals and is based on the mapping between the heterogeneous dataset's metadata delivered by some OD portal platforms (e.g. CKAN, Socrata,...), and the W3C-DCAT‡ vocabulary aimed at facilitating interoperability between data catalogs published on the Web. Due to these features, we deemed it feasible to map Figshare metadata into W3C-DCAT, and afterward assessing its quality. The platform, implements fifteen objective quality metrics to assess the compliance of ingested datasets metadata with respect to DCAT. The metrics exposed by the platform concern three quality dimensions: (i) Existence (i.e., the extent to which important metadata keys for a dataset are provided); (ii) Conformance (i.e., the extent to which metadata values adhere to a certain format; (iii) Open Data (i.e., the extent to which the specified format and license information may classify a dataset as open). We integrated our gathering usage code with the 'Open Data Portal Watch' source code, and extended it to operate with Figshare metadata as well as to elaborate and produce analytic and reporting. Essentially the whole platform follows a simplified ELT (Extract Load Transform) approach. It extracts metadata from the portals, and loads them as a whole as JSON§ objects on the internal database, finally gathers and transforms the required information chunks to assess quality and usage behavior, storing the results on the internal database to allow further analytics.

   The initial hypothesis of quality assessment was clearly connected to the FAIR tool. However, after performing several tests on datasets belonging

---

*actually referred to as 'article' in Figshare documentation
†https://github.com/sebneu/portalwatch
‡https://www.w3.org/TR/vocab-dcat/
§https://www.json.org/json-en.html

to different Open Data portals, included Figshare, we realized that the waiting times for each evaluation carried out (on the 22 metric available) by the FAIR tool were never less than 5 minutes with peaks of 30 minutes or more. The reason for this delay is probably due to the fact that for each single dataset, a series of multiple calls to remote sites[*] are necessary to check if a given element present in the metadata of the dataset conforms to the underlying principle. For this reason, the decision of working on a huge pool of data (all Figshare resources) led to the analysis of a different approach for the quality assessment. In any case the selection of quality parameters in this article is largely aligned with FAIR metrics.

Data Analysis The data was processed and bivariate descriptive statistics were elaborated for each of the categories under analysis (Quality, Views and Downloads per Research Field). In order to observe further relationships between the variables, Spearman correlation between Views, Downloads and Quality has been carried out. This type of correlation, was preferred as non-parametrical analysis given the possibility of unusual distribution. Finally, a k-means cluster analysis was undertaken with the aim of observing consistent patterns across research field in views and downloads per quality. Taking into consideration the observed values of several variables, the technique of cluster analysis aim at grouping similar observations. As such, it can be considered a multivariate statistical technique, which adopts a number of different methods. In fact, according to the types of algorithms used, we get unsupervised and supervised methods. In our case, we applied the simplest technique, unsupervised k-means clustering. This method is based on vector quantification, for it partitions n observations (in our case, the n of items extracted from Figshare) in k clusters, in which each observation belongs to the cluster with the nearest mean. The algorithm launches a number of iterations which minimizes distances between individual points in a cluster and the cluster center. To determine which variables were the most effective for distinguishing cluster, it was adopted the Analysis of Variance computed per variable and its resultant variance table. In this regard, F-statistic, p-value, Model Sum of Squares and Degrees of Freedom, Error Sum of Squares, and Degrees of Freedom, were the variance statistics for

---

[*]https://github.com/FAIRMetrics/Metrics/tree/master/MaturityIndicators/Gen2

clustering included in our analysis. The variables used to generate the clusters were: Views, Downloads and ORD Quality, adopting the medians as measures. The categorical variable adopted as concept to group the observations was the research fields. Therefore the clusters yielded further information over the trends of usage across disciplines.

Results

RQ1: Which are the forms of publication and usage of ORD, considering the scientific fields of FigShare?

The initial operation to address RQ1 was to explore the forms of publication, usually the starting activity involving a researcher when dealing with ORD. The Table 1 shows the overall sample characteristics, in terms of overall number of resources per categories. Almost two third

Table 1. Figshare 21 Scientific Fields

Category	Sub-categories	Total Items	% Items	Figure	Dataset	Journal	Other
Engineering	137	34,671	0.6%	2,627	13,510	9,385	9,149
Physics	45	351,615	5.6%	36,768	75,485	206,434	32,928
Psychology	29	202,307	3.2%	67,954	81,107	47,487	5,759
Social Science	48	280,082	4.5%	87,674	85,843	98,628	7,937
Uncategorised	1	64,197	1.0%	11,501	5,950	19,542	27,204
Earth and Environmental Sciences	90	864,774	13.8%	275,956	287,323	272,819	28,676
Chemistry	70	1,269,242	20.3%	344,628	351,763	526,812	46,039
Meta Science	2	458,185	7.3%	176,721	186,315	76,228	18,921
Astronomy, Astrophysics, Space Science	22	159,242	2.5%	3,626	7,218	1,780	
Biological Sciences	114	1,972,878	31.5%	865,653	567,931	482,718	56,576
Humanities	163	8,004	0.1%	958	2,327	2,385	2,334
Information and Computing Sciences	73	422,733	6.8%	251,430	80,156	74,100	17,047
Mathematics	44	63,354	1.0%	20,041	24,756	15,640	2,917
Health Sciences	144	61,507	1.0%	13,870	28,952	12,540	6,145
Studies in Creative Arts and Writing	32	2,760	0.0%	236	525	585	1,414
Technology	62	8,906	0.1%	1,241	6,848	189	628
Built Environment and Design	36	7,546	0.1%	238	430	2,941	3,937
Commerce, Management, Tourism and Services	54	5,609	0.1%	93	895	3,554	1,067
Studies in Human Society	69	9,396	0.2%	205	2,765	4,572	1,854
Language, Communication and Culture	79	6,351	0.1%	2,885	236	1,817	1,413
Agricultural and Veterinary Sciences	66	9,068	0.1%	3,632	4,159	478	799
		6,262,427		2,310,929	1,810,902	1,866,072	274,524

of resources come from three research area, *Biological Science* (31.5%), *Chemistry* (20.3%) and *Earth and Environmental Sciences* (13.8%). Another 25% of resources belong to *Meta Sciences* (7.3%), *Information and Computing Sciences* (6,8%), *Physics* (5.6%), *Social Sciences* (4.5 %). The remaining 10% is contributed by the other 14 categories. More than 95% of resources belong to one of three types, i.e. Figures (37%), Datasets (29%) and Journal contributions (30%). Apart of the huge productivity of three main scientific fields, it seems that all of them

contribute in a balanced way in terms of types of resources. The within-category comparison will uncover further, diversified dynamics of ORD production.

The Figure 1 shows the situation within the scientific fields. As we
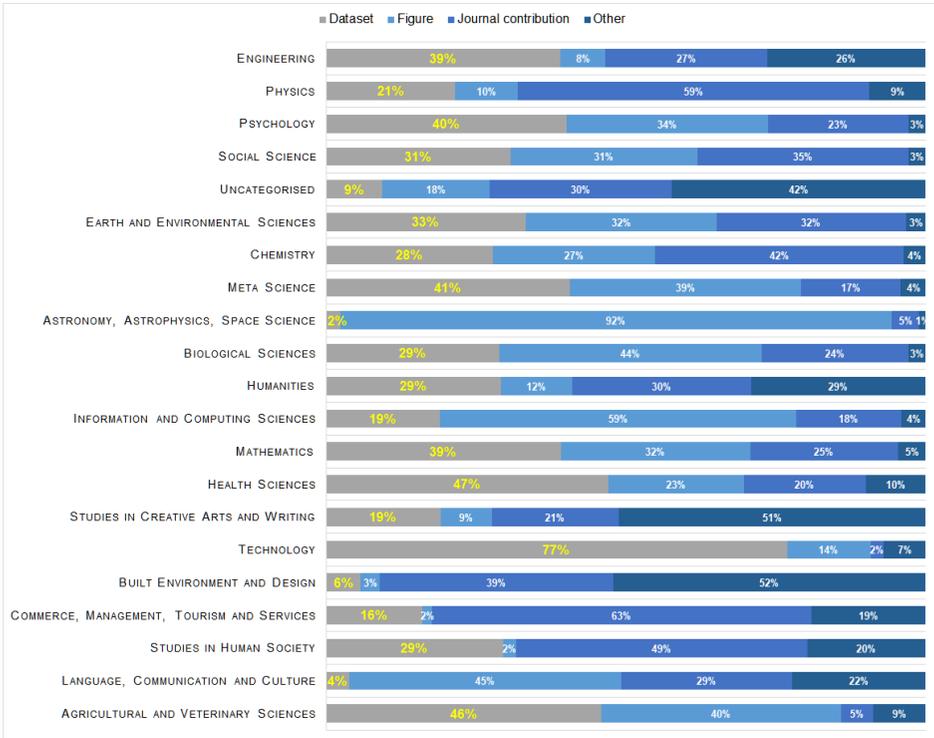

Figure 1. Percentage of Figshare resources per category, with respect types 'Dataset', 'Figure', 'Journal contribution' (and Others)

can observe, most important contributors come from natural sciences (*Biological Science*, *Chemistry* and *Earth and Environmental Sciences*). These three disciplines publish circa a third of their resources as ORD; however, while *Biological Sciences* tend to contribute with more figures (44%) *Chemistry* shows a more relevant concentration of Journal contributions (42%). As for the disciplines with a prevalent contribution in terms of ORD (like *Technology*, 77%; *Health Sciences*, 47%; *Agricultural and Veterinary Sciences*, 46% ; *Meta Science*, 41%; *Psychology*, 40%; and *Engineering*, 39%) they are part of the tiny 10% for overall Figshare resources. However, the relative comparison shows a culture of ORD

publication amongst those publishing, namely, the fact that the researchers arriving to use Figshare are informed or consider relevant the ORD publication. The very little number of ORD published by scientific fields as *Language, Communication and Culture* (4%) and *Built, environment and Design* (6%) should also be noticed. These might be linked to the diversified type of research units in these sectors (little need of numeric, machine readable datasets against complex objects like representations, graphics, digitized cultural heritage objects, etc.). All other 11 scientific fields fall into the detected percentages of 20-35% of ORD publication, including the 'big three' (i.e. *Biological Science*, *Chemistry* and *Earth and Environmental Sciences*).

To continue exploring the researchers' behavior in relation to ORD, we also considered the forms of usage, operationalised as views and downloads. The usage of second-hand ORD is a further operation a researcher undertakes, when aiming at collaborating with other researchers or informing own processes of ORD creation and publishing. We hence considered both the descriptive bivariate statistics showing the relationship between scientific fields per Views and Downloads. To further in the study of this relationship, the Spearman correlation was also considered. Finally, in order to see if there were possible aggregation of trends of usage, the cluster analysis was implemented.

Views and Downloads values gathered from Figshare provided further information about the usage going beyond the single author, to a more social parameter. This is, in fact, a superficial interest (view) to a more relevant form of interest in which there is at least an intention of consultation with eventual impact on the second researcher (user) practices. In the following, we comment separately the descriptive stats of views and downloads.

Views The Figure 2 presents the percentages of views relating the three types of resources (Dataset, Figure and Journal contribution - and others). As we might observe, the views of datasets are consistent with the trends observed in the publication of ORD against the other resources. As a matter of fact, *Technology*, which is one of the scientific fields with more ORD publication, even though it contributes an insignificant number of overall resources to the Figshare in comparison to other disciplines, accounts for the most viewed ORD (69%); followed by *Agricultural and Veterinary Sciences* (55% of views against the views of all other

resources in this disciplinary field); *Meta Science* (49%); *Psychology* and *Mathematics* (47%). The 'big three' are viewed also consistently with their patterns of publication can be collocated between 36-43%. The two scientific fields with the least publication of ORD are also consistent: 2-3% of views in the case of *Language, Communication and Culture* and *Built Environment and Design*. However, there are some cases in which the Views are far more than the expected for the number of resources published. This is the case of *Engineering* (9% of views on ORD against 39% of published ORD); or *Health Sciences* (26% against 47%). So it looks that in these disciplines there are far less readers/direct users than authors.
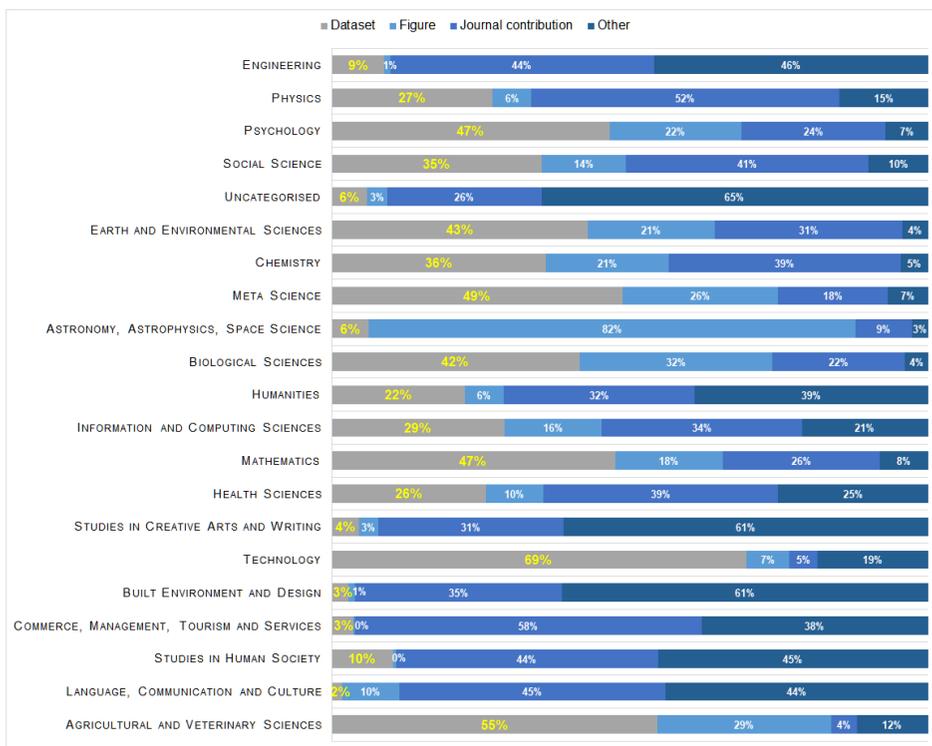

Figure 2. Percentage of Views with respect three types 'Dataset', 'Figure', 'Journal contribution' (and Others)

The Figure 3 shows the views distributions of frequencies combined with the Figshare categories. In all cases, the curves show heavy-tailed

distributions with very few datasets with a high frequency of use, and most of them with very low frequencies of usage.
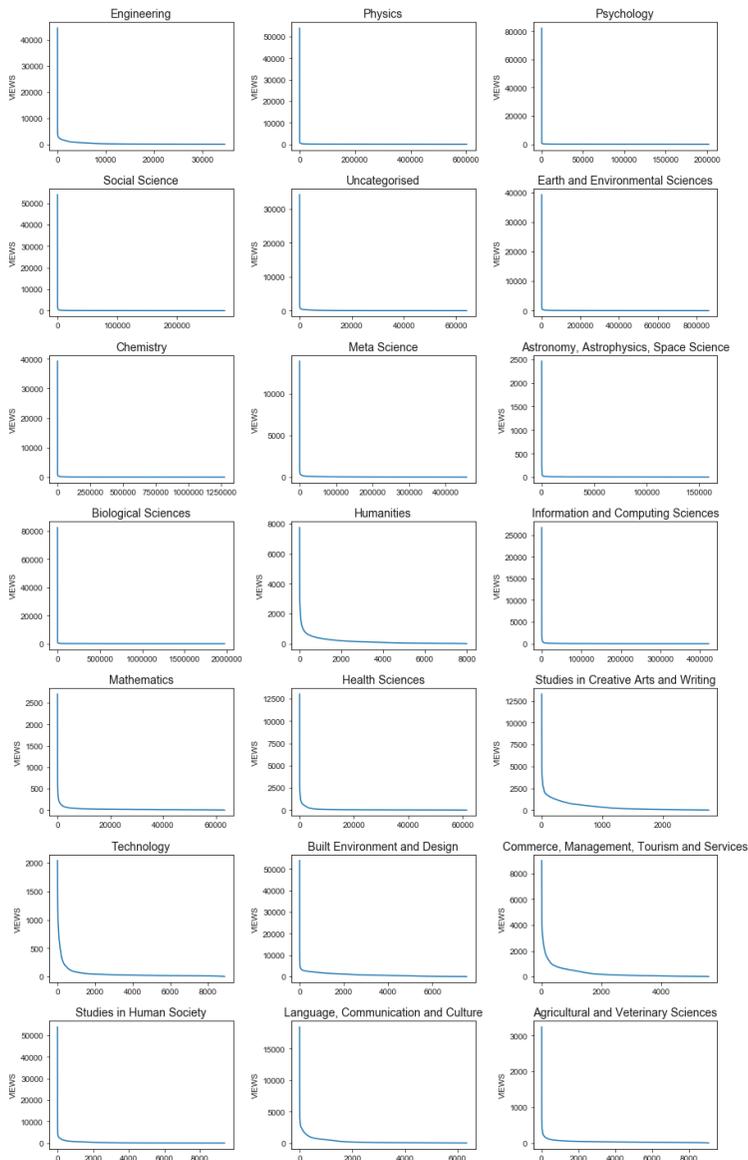

Figure 3. Views distribution for categories

Downloads When coming to the comparisons between ORD downloads against other resources in the Figshare platform, one notices again overall consistency between scientific fields that are most productive in terms of ORD and the downloaded materials. The Figure 4 shows the percentages of downloads relating the three types of resources (Dataset, Figure and Journal contribution - and others).
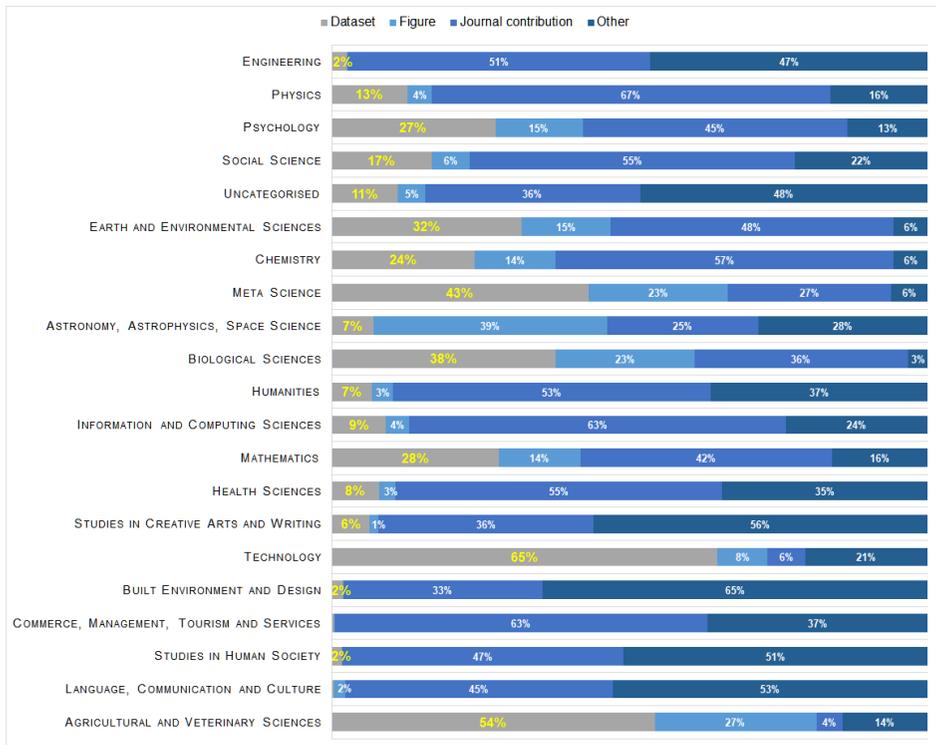

Figure 4. Percentage of Downloads with respect three types 'Dataset, 'Figure', 'Journal contribution' (and Others)

As a matter of fact, *Technology* shows a trend of 65% of downloaded ORD against 77% of published ORD; *Agricultural and Veterinary Sciences*, 54% against 46%; *Meta Science* 43% against 41%. The 'big three' contributing to about the 66% of resources in the Figshare platform have a balanced situation where the ORD are circa one third of the produced resources and the downloads are about one third of the downloads (*Biological Science*, 38% against 29%; *Chemistry*, 24% against 28%; *Earth and Environmental Sciences*, 32% against 33%). Also

scientific fields contributing the least with ORD show a balance between the downloaded against the produced (*Language and Communication*, 2% against 4% and *Built Environment and Design*, 2% against 6%). It is also possible to see slight differences in some few cases where the produced represents relatively less than the downloaded (*Biological Sciences* or *Agricultural Sciences*). In all other cases the productivity of ORD seems to go faster than the usage in terms of downloads. A particular situation is represented by scientific fields where there is little activity of consultation, like *Health Sciences* (8% of downloaded ORD against 47% of produced ORD), *Engineering* (2% against 39%) and *Psychology* (15% against 40%)

Moreover, the analysis of Downloads threw a similar situation with regard to the Views in terms of frequencies of usage. As it is shown in Figure 5 the distributions were highly skewed, what can be interpreted as little intention of usage (expressed as downloads) for most resources and massive usage of a handful of ORD.

These results are confirmed by examining the descriptive statistics reported in Table 2, which supplies more insights on the usage trend comparing Views and Downloads. In particular, it is noted that the median values of the downloads are always lower than the views, indicating that resources are generally more seen than downloaded. However, for some categories, it appears that the total downloads exceed those of the views. This fact is shown by Figure 6 that synthesizes the previous observations, reporting the percentage of views, downloads, and the number of items for each category, wrapping up the ongoing situation of production of research resources against its usage. From Figure 6 is it clear that in some cases (e.g *Engineering*, *Information and computing science*) the overall number of downloads is greater than the views. This situation, is not inconsistent, as reported to us by the Figshare support desk: "The case you are describing is not out of the ordinary. In some cases the number of download can be higher than the number of views, and here are 3 main reasons to support that: i) For items where there are multiple files attached, there are multiple download links(one for each file) and we count 1 download for each file download. So you will have 1 view and 99 downloads, for example for the item you have mentioned; ii) If items are harvested through API calls, there will be downloads counted, but zero views; iii) Bots activity."

From Figure 6 we can observe another interesting fact on the disparity between the number of resources published and their use. The categories
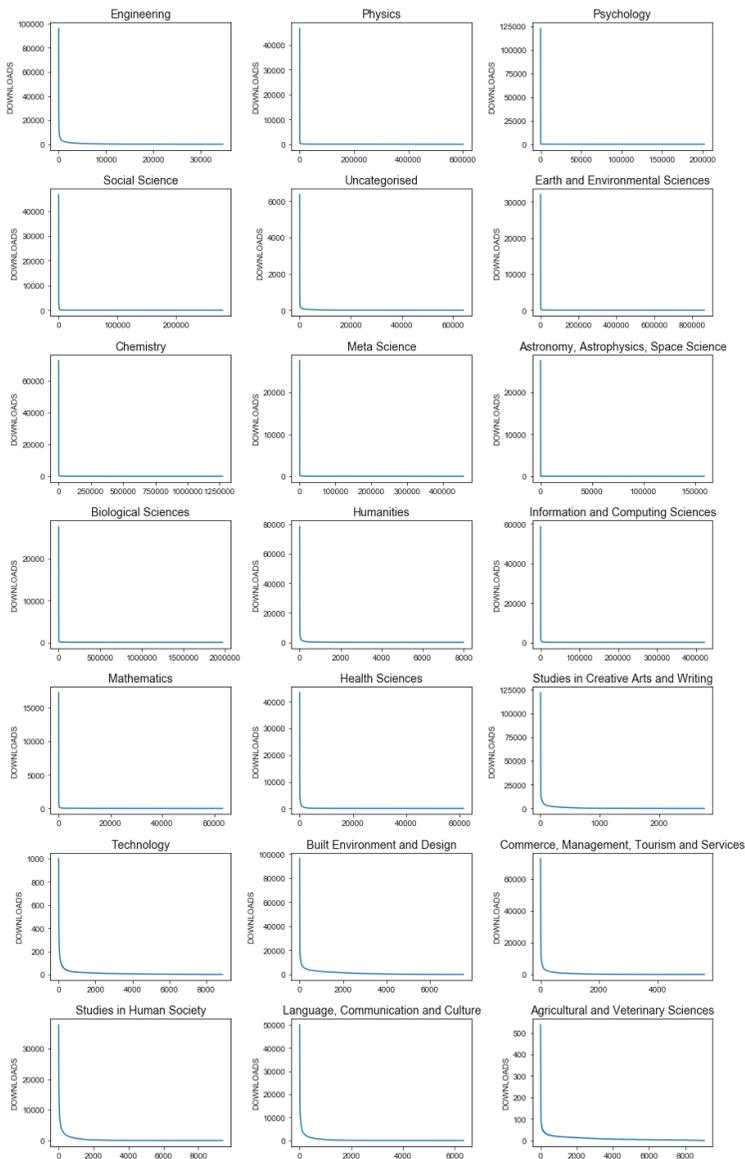
Figure 5. Downloads distribution for categories

that publish the most are also the ones that, proportionately, are visited less in percentage terms. The 'big three' with the presence of two-thirds of the total resources are seen just under 60% but downloaded about

Table 2. Descriptive statistics for Views and Downloads

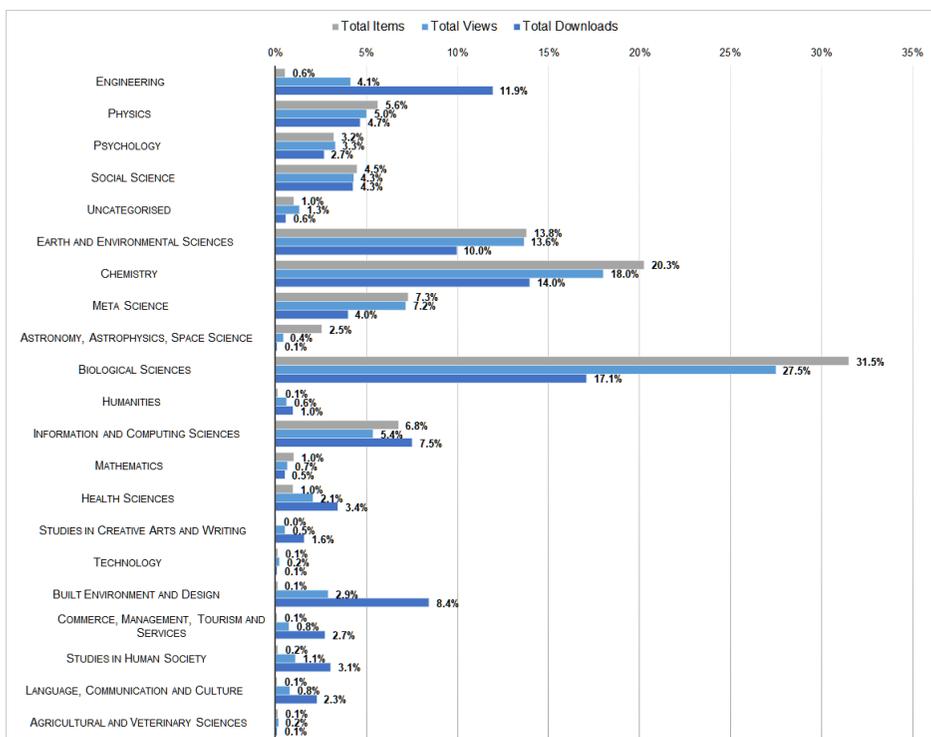| Category	Views				Downl.			
	Mean	Stdd	Median	Max.	Mean	Stdd	Median	Max.
Engineering	258	584	52	44,488	359	1,354	19	96,210
Physics	31	88	16	19,819	14	104	5	27,623
Psychology	35	284	19	82,252	14	310	4	122,542
Social Science	33	171	15	54,027	16	217	4	46,743
Uncategorised	45	183	7	34,228	9	43	2	6,360
Earth and Environmental Sciences	34	94	18	39,290	12	87	4	32,030
Chemistry	31	74	17	39,291	11	73	5	51,774
Meta Science	34	83	19	13,917	9	52	4	27,625
Astronomy, Astrophysics, Space Science	6	14	5	2,466	1	69	0	27,623
Biological Sciences	30	104	17	82,408	9	51	4	27,731
Humanities	172	310	79	7,758	128	1,148	16	78,448
Information and Computing Sciences	28	154	7	26,750	19	248	1	58,715
Mathematics	23	52	13	2,707	8	105	3	17,263
Health Sciences	73	218	26	13,048	58	476	5	43,419
Studies in Creative Arts and Writing	411	622	164	13,289	604	3,296	52	121,995
Technology	53	119	24	2,046	11	31	4	1,000
Built Environment and Design	835	1,186	580	54,060	1,162	2,431	477	96,210
Commerce, Management, Tourism and Services	292	523	105	9,025	510	2,140	83	72,952
Studies in Human Society	260	854	39	54,022	339	1,401	9	37,830
Language, Communication and Culture	273	640	64	18,476	372	1,644	19	50,146
Agricultural and Veterinary Sciences	39	59	28	3,247	9	17	6	538

Figure 6. Percentage of Number of items, views and downloads per Figshare categories

40% of the total. Categories with medium-small numbers of resources, such as *Meta science*, *Physics*, *Social science* and *Psychology*, have a percentage of views almost identical to the number of resources and a download percentage just below. There is then the exception of *Astronomy, Astrophysics, Space Sciences*, which with a 2.5% of resources is visited by 0.4% users and downloaded just by a 0.1%. Compared to those, *Information and Computing sciences* is in contrast with the number of downloads (7.5%) higher than both the published resources (6.8%) and the number of visits (5.4%). The most surprising aspect concerns the remaining categories which cover just over 4% of the resources. All of these, apart from *Uncategorized* and *Mathematics*, are viewed over 13% of times and downloaded 34% of the time. Among them, *Built environment and design* and *Engineering* stand out. On the other hand, for these two categories, Figure 2 and 4 tell us that most of the views and downloads are not related to datasets, which are present to a large extent (39% in *Engineering*), but to other types of resources. These observations highlight how the effort to publish research resources is not always compensated by proportional feedback. A similar situation is observed by Berends et al. [46] (p.27) analyzing the data published by the European Data Portal and their reuse, which indicate that there is a "mismatch between available data sets and re-used data". To overcome this discrepancy, the authors suggest public administrations "to develop a publication strategy which is more aligned to the needs of data users". However, in the case of Figshare, and of other scientific portals, that suggestion finds difficult application, as unlike what happens in governmental portals, in which the managers follow centralized publication policies, the contributions in the repositories are the choice and responsibility of individual researchers and research groups. At most, they can be recommended to publish contributions (ORD or other types of scientific resources) of higher scientific quality, or at least to describe these contributions in clearer terms, to facilitate their access and visibility. We will see in the next section, if and how the quality of scientific metadata is related to their use.

The relationship between Views and Downloads The descriptive statistics led to consider the relationships between variables in spite of the skewed distributions. The application of the Spearman correlation of downloads w.r.t. Views yield a significant result: $\rho = 0.7$, $p-value =$

0.0. In this light, we considered the possibility of analysing the groups formation under this overall figure. The relationship between Views and Downloads supported the formation of three clusters, showing some differences in the patterns of usage (Views and Downloads) by scientific field (as defined by Figshare, with all the precautions). The figure (Figure 7) shows the clustering with the relative trend models of correlation of values within the cluster, later explained in the tables.
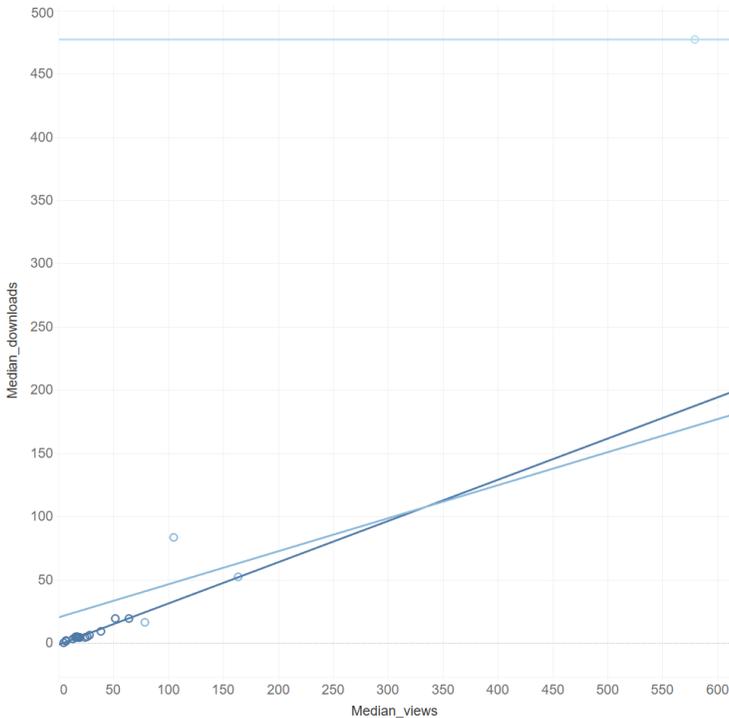

Figure 7. Cluster Analysis: Views and Downloads

Table 3 shows the Figshare scientific fields grouped within the 3 clusters identified; Table 5 shows the Cluster variance statistics and Table 4 shows each cluster model (based on the correlation of quantitative variables "Views" and "Downloads").

As we observe the only significant model (predicting inner-cluster components correlation) relates the Cluster 1. This is composed by most scientific fields, namely, seventeen of them. This first cluster groups scientific fields that show a very limited number of views and downloads

Table 3. Figshare Categories grouped by three identified clusters

Cluster	Category	Median Views	Median Downloads
Cluster 1	Language, Communication & Culture	64	19
	Engineering	52	19
	Studies in Human Society	39	9
	Agricultural and Vet Sciences	28	6
	Health Sciences	26	5
	Technology	24	4
	Psychology	19	4
	Meta science	19	4
	Earth and Environmental Sciences	18	4
	Chemistry	17	5
	Biological Sciences	17	4
	Physics	16	5
	Social Sciences	15	4
	Mathematics	13	3
	Uncategorised	7	2
	Information and computing sciences	7	1
	Astronomy, Astrophysics, Space Sciences	5	0
Cluster 2	Studies in Creative Arts and Writing	164	52
	Commerce, Management, Tourism	105	83
	Humanities	79	16
Cluster 3	Built environment and Design	579,5	477

Table 4. The linear trend model was calculated for median_downloads given median_views. The model can be significant at $p <= 0.05$. The Clusters factor can be significant at $p <= 0.05$. Discordance analysis (ANOVA) df = 3, SSE = 5,110.13, MSE = 1,703.38, F = 13.4606, $p-value < ..0001214$

Model Formula:	**Cluster***
	(Views + intercept)
Number of modeled observations:	21
Number of filtered observations:	0
Degrees of freedom of model:	5
Degrees of residual freedom (DF):	16
SSE (sum of squares of error):	2,024.73
MSE (mean square error):	126,546
R square:	0.990508
Standard Error (StdErr):	11,2492
p-value (cutoff):	<0.0001

Table 5. Individual trend lines

	Line		Coefficients				
Cluster	p-value	DF	Term	Value	StdErr	T value	P value
Cluster 1	<0,0001	12	Median views	0.326059	0.0241539	13,4992	<0.0001
			intercept	-1.63876	0.661586	-2.47702	0.0256418
Cluster 2	0.779631	3	Median views	0.261202	0.724199	0.360677	0.779631
			intercept	20.0339	87.8661	0.228005	0.857287
Cluster 3	N/A	0[*]	Median views	0	-	-	-
			intercept	477	-	-	-

[*] The trend line model has zero residual degrees of freedom (no information to estimate the model)

(medians, excluding outliers), and does not relate to a specific area of knowledge (like humanities or natural sciences). One important notice is that this cluster includes the scientific fields with highest number of ORDs. The second cluster is composed by far less scientific fields (*Studies in Creative Arts and Writing*; *Commerce, Management, Tourism*; *Humanities*) which are, to some extent, related fields of knowledge. In this second cluster, the relationship of views and downloads are slightly more balanced with regard to the first cluster, even though these scientific fields contribute with a rather insignificant number of cases to the sample of ORD published on Figshare. Notwithstanding, in this group, the researchers appear to download more consistently what they decide to view. Lastly, the third cluster composed by only one category (*Built environment and Design*) present the highest median and more balanced relationship between views and downloads. In this case, the curious finding is the little number of published ORD, but the frequent interest and motivation to re-use them (at least as information downloaded) by other researchers from the sector.

RQ2: Which is the metadata quality of ORD in Figshare?

To address RQ2, we apply the metadata quality assessment tool to all the 21 Fighsare disciplines. As can be seen from the histograms on the distribution of quality values (Figure 8), overall metadata quality (obtained by aggregating on the different metrics) shows a distribution with the mean values in the range $0.5 - 0.6$. However, from Figure 8, it can be noted that, despite the common platform, there are some differences between categories. For a few of them, the average and median values are around 0.44 (e.g. *Humanities*, *Studies*
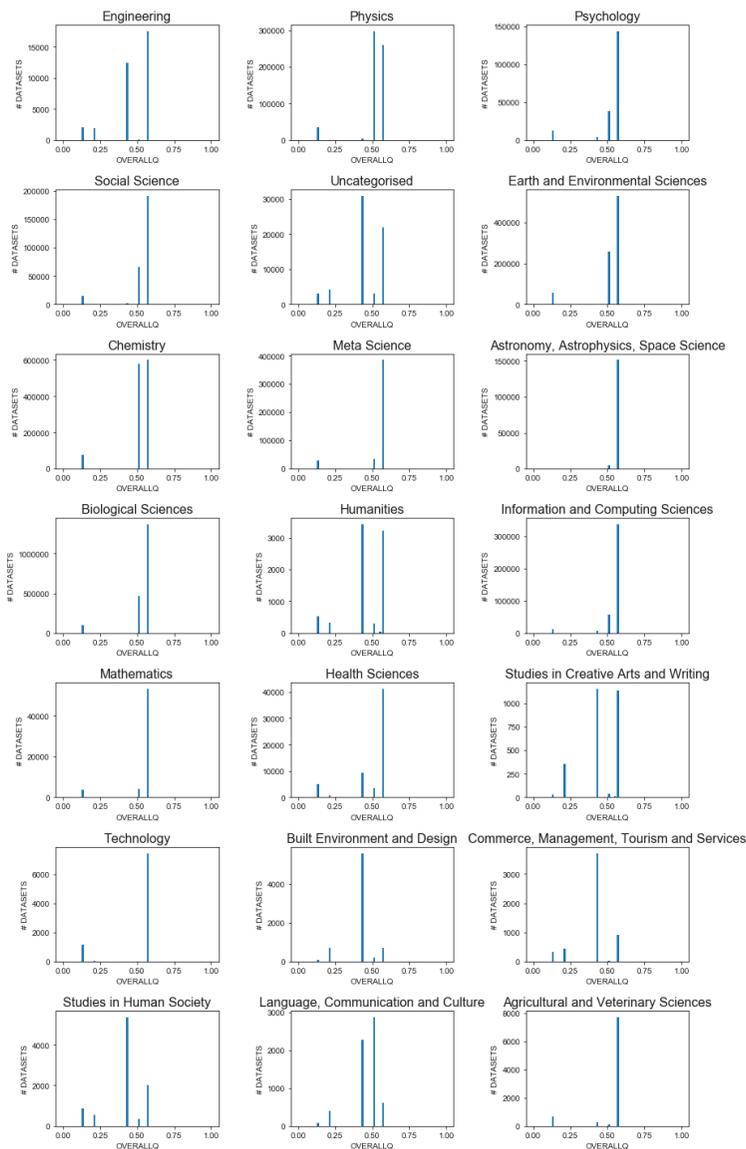
Figure 8. Overall quality distribution for Figshare categories

in Human Society, *Uncategorized*), while the majority of others, such as *Biological Sciences* and *Mathematics*, have median values around 0.57. The reason for this difference lies in the different content of some metadata field, which value is inserted by datasets providers,

not in the figshare metadata schema. For example, by examining the metadata associated with some *Uncategorized* and *Biological Sciences* resources, we noticed that license information, although being present in both metadata, varies. For instance, a resource of *Biological Sciences* has the string 'https://creativecommons.org/licenses/by/4.0/' as the license URL, which is recognized to be an open format. This fact contributes to increasing the total quality value for the resource. By contrast, an *Uncategorized* resource has the license URL set to 'http://rightsstatements.org/vocab/InC/1.0/', therefore not acknowledged as an open format. This fact lower the quality score of the associated resource.

RQ3: Does the metadata quality influence the forms of ORD usage in Figshare?

To verify the RQ3 we explored the correlation between number of views and metadata quality, considering that the trend of the first does not follow a normal distribution (see Figure 3), thus excluding the Pearson test, we resorted to Spearman's rho non-parametric test.

Applying Spearman indifferently to all the collected data we obtained a rho value $\rho = -0.09$ with $p = 0$. This rho indicates a very small negative correlation even if significant. We then analyzed the behavior of rho on the individual categories, and the observable results plotted in Figure 9, testify that in just four cases the rho values are at most, close to 0.4 (i.e. *Information and computing sciences*, *Engineering*, *Language, Communication and Culture*, *Studies in Creative Arts and Writing*); therefore a correlation value generally considered medium-low[67]. However in most cases significant values are far lower, and in one case (*Technology*) not significant at all. Furthermore, for the 20 categories with a significant rho, the sign of the correlation is mainly negative (15 out of 20). These values indicate that the quality of metadata (measured with our approach) is not positively related to its use. Rather, albeit with a low rho value, the negative sign suggests that users are more interested in viewing resources with poorer metadata quality, thus apparently contradicting the assumption that good quality metadata is a prerequisite for access and the reuse of open data[15,16].

Since the Spearman rho showed a significant value, the cluster analysis was carried out. Applied to Views/Quality and Downloads/Quality it
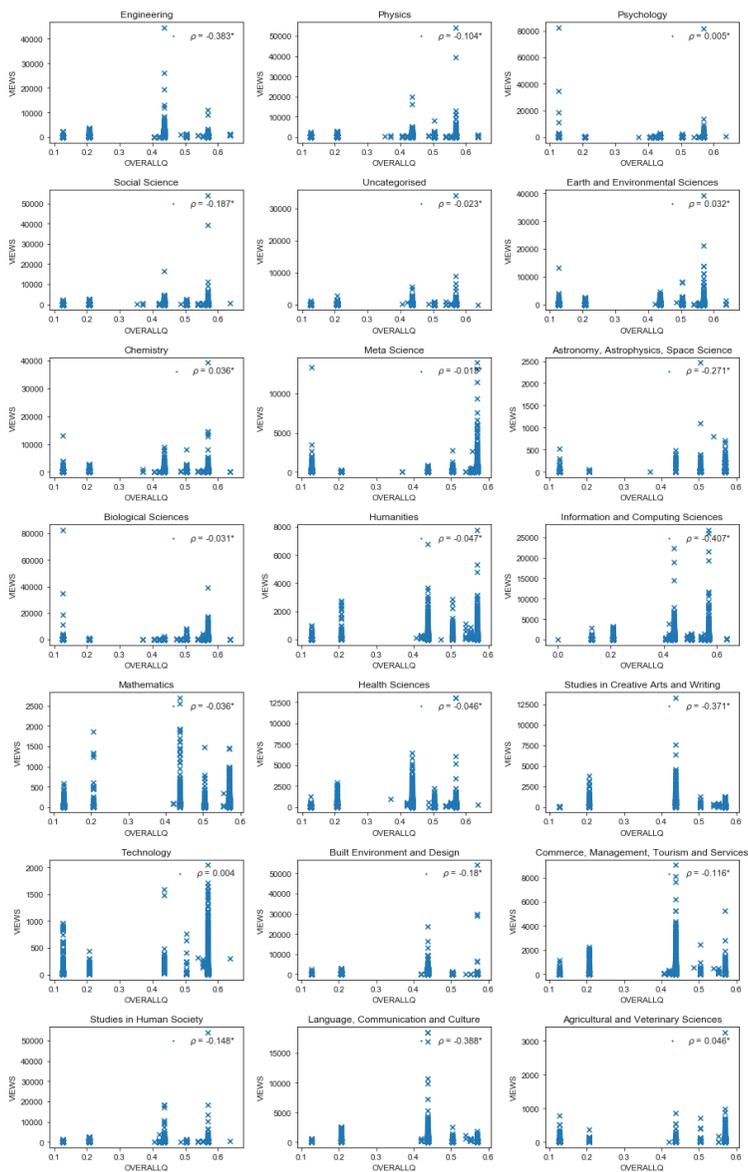
Figure 9. Correlation between quality and number of views. Spearman's ρ values with $p < 0.05$ postfixed with *

yielded significant p-values at the $< .0001$ cutoff value, with two clusters detected (View/Quality, $SSE = 7,672.22$, $R2 = 0.975297$, $df = 4$, $p-$

$value = .0001$; Download/Quality, $SSE = 2,693.76$, $R2 = 0.987371$, $df = 4, p - value = .0001$) No significant correlation was found within each of the clusters (meaning that the relationship between elements was not relevant) but the separation of two groups of "social activity" relating the quality of the clusters was relevant. The patterns observed pointed out at a first cluster (same scientific fields seen in Table 3) connecting relatively low quality objects (0.4) with low levels of usage (between 79 and 164 median views and $16 - 83$ median downloads per ORD); and a second cluster with paradoxically better levels of quality $(0.5 - 0.6)$ connected to even lower levels of usage ($7 - 64$ views and $2 - 19$ median downloads).

Discussion

In this section we will discuss the three research questions on the light of the findings. The first finding relating the RQ1 highlights a situation that is diversified, between the initial publication and the further "social" activity by direct users of ORD. Indeed, the comparison between ORD and other types of resources to Figshare like figures and journal contributions showed a general balanced situation of ORD creation and usage. Figures and articles are more typically shared as main or subsidiary material in some scientific fields, while others are advancing at a fast pace in the creation of ORD as part of the research routines. Indeed there are disciplines where data-driven practices are increasing more than in others, but this is also due to the characteristics of study objects too. Another issue is that the Open Access policies, started early before the Open Data policies, could have already created a base of technical skills amongst researchers to take them to publish other resources more than datasets. Even if the approaches amongst disciplines and regions might differ, there are ongoing trends relating OA publication[30,68]. As a result, many researchers are already used to share their pre-prints or post-prints of published research articles and the subsidiary resources (like figures) to Institutional Respositories, and mostly, to academic social networks like ResearchGate[40,68,69]. One must consider two issues when analysing this finding. Firstly, the ORD publication are a rather recent practice, particularly in the way of a requested and compulsory activity when a research project has got fund. The expected impact could be of having less ORD relating to other types of resources. However, Figshare

was originally born as open data portal and the research community approaches it mostly to upload ORD. The figures are mostly balanced, both for those scientific fields that cover the two-thirds of uploads to Figshare (*Biological Sciences*, *Chemistry*, *Environmental Sciences*) as well as for a number of disciplines contributing with least resources but with clear patterns of Open Data creation (like *Technology*, *Agricultural Sciences*, *Health Science*, *Meta Science*, *Psychology* or *Engineering*). This is an evidence of the fact that the researchers are slowly embracing the ORD publication, with differences due to the inner characteristics of the study objects, as it had been early explained in[7]. In any case, one cannot expect a linear embracing of Open Data publication and sharing practices. As explained in[17], these are heavily connected with cultural and political contexts of data friction, or the phenomenon of discontinuity and disparities in the circulation of data. In our extractive and quantative approach to the research object, we could only uncover the existing practices. To deepen on the causes of data friction, ethnographic and narrative methods should be considered in order to unveil the politics of data friction across disciplines.

As for the connections between publication and usage of ORD (RQ1), the deeper look at the direct users' activity once the ORD has been published, seems to show a rather different panorama. As it comes out from the raw numbers on views and downloads in Table 2, across all research fields, the majority of the published datasets is used just marginally. Our results confirm the concerns of some CDOs reported in[70], according to which "We counted the clicks and we saw that these portals just weren't being used". A common pattern for all scientific disciplines is the highly negative distributions (left skewed). This implies that there are ORD at the third quartile which show a peak (both for views and downloads). Instead the first quartile tend to show values nearer to zero, meaning virtually no views or downloads for an ORD. One should still consider that there are disparities between cases, and not all highly viewed resources present a relevant number of downloads. The disparities between the median and the third quartile (top values) are extremely unbalanced in some cases like *Biological Sciences* (17; 82,408); or *Social Sciences* (15; 54,027) or *Earth and Environmental Sciences* (18; 39,290). This situation is similar to other scientific fields, but with a slightly more balanced situation, with less distance between the median

and the top levels. For example, the domains of *Studies on Creative Arts and Writing* (164; 13,289); *Commerce, Management, Tourism and Services* (105; 9,025) and *Technology* (24; 2,046) are representatives of this trend. In the case of *Built Environment and Design* (580; 54,060) it is possible to see that the least viewed resources yet capture a modest attention. This point outs different researchers activity around ORD by the research fields defined by Figshare. It might be also considered that the research fields with less ORD are those showing more balanced trends. In the case of most productive categories, the attention seems to be focused over specific ORD. The cluster analysis reinforced this picture, by showing that most research fields grouped are characterized by very low usage (in terms of views and downloads), with very few exceptions; and high concentration of attention over seemingly specific datasets. In this regard it is extremely important to consider that the research fields as defined by Figshare could encompass differences between self-defined research fields, classification by scientific databases like ISI-Web of Science[*] or Scopus[†] or canonical nomenclatures such as the UNESCO-SKOS (international standard nomenclature for fields of science and technology[‡]). Therefore, the findings hereby characterised by research field must be considered with due parsimony. Although the choice of metrics we adopted can influence the extent of the assessments on the use of the dataset, they provide significant indicators to ORD providers to understand if the datasets published on the portals they manage attract the interest of the users. This fact resonates with what reported, in the previously cited survey, by one CDO "We look at the total number of datasets that are out there, what we are offering up. We count visit clicks, and lastly, we look at how many downloads are actually being done off the open data portal"[70].

   As to RQ2, our findings show a seemingly appropriate metadata quality for almost all Figshare ORD, with an overall median of 0.57, with few exceptions, identical for all the disciplines. These results can certainly be ascribed to the metadata editor provided by Figshare, which guides scientists in the publication of ORD metadata, reducing, at least in part,

---

[*]www.webofknowledge.com
[†]www.scopus.com
[‡]https://skos.um.es/unesco6/

some of those Data friction factors which lowering the metadata quality, hinder the reuse of ORD[17,21].

When coming to the RQ3 taking into consideration usage vs. quality, our results did not support the idea that the highest quality encompasses more attention (views) and eventual re-use (downloads). Our findings point out a mostly random behavior, where the users might be pushed by other factors far from clear metadata, open licences to re-use objects, etc. Moreover, the inverted rho values underlined a situation where some ORD with insufficient quality parameters are preferred to better quality ORD. We found significance in clustering when using quality parameters, but there was sparsity within the second cluster, which model also showed a negative correlation, a result that underpins our conclusion that other factors weight over the ORD direct usage beyond the metadata quality. Moreover, this conclusion would be further supported acknowledge by the socio-critical lens of[17], which highlight the complex nature of researchers' (and other stakeholders) behaviors in circulating the ORD, producing voluntary or involuntary data friction. There are two important considerations that could be done over the causes and forms which data friction might assume. The first related researchers' data literacy, which, as we considered in the background, deserves attention. The skills' gap could hinder researchers, both by their complete lack of awareness or their partial and uncompleted awareness of ORD. The second, more complex motivation relates the phenomenon early discussed by Robert Merton in his foundational work on the normative structure of science[71]. In this regard, there are hidden rules connected to the research cultures across fields which pervade and guide researchers' attention, decision over topics and methodologies. These cultural factors could be influencing researchers at the time of focusing their attention to most influencing researchers in their fields (ORD with highest views and downloads) relating to peripheral researchers or newcomers. These behaviour is also supported by the theory of Legitimation of peripheral participation[72], where newcomers are progressively accepted while showing similar behaviors and codes relating to central components within professional communities (or communities of practice). The element requires a different approach of research (like a qualitative set of interviews) over the basis of objective parameters extracted from Figshare, to shed light over these assumptions. By now, the emergent phenomenon is the

huge concentration of views and downloads over specific ORD beyond its quality. Moreover, defining quality through automated procedures, namely, metadata quality parameters, could be tricky. Quality can assume several perspectives according to the positioning of stakeholders[73]. In this regard, even if metadata is in place, there might be other values connected to the subjective perception of quality by ORD second-hand users.

Conclusion

In this paper, we studied basic parameters of usage of ORD, in order to further the evidence on the advancement on Open Data as essential piece of the Open Science. Our research design was based on an exploratory study where we analyzed the objects extracted from open data portals in terms of quantity, type, research field, metadata quality and usage. We formulated three research questions addressing data collection and analysis, in connection with the study aims above introduced. Our research questions focused on: 1- The forms of publication of ORD relating other type of resources (Figures, Journal contributions, etc.), and the forms of usage of ORD, in terms of views and downloads, all across research fields; 2 - The metadata quality of Figshare ORD; 3 - The relationship between metadata quality of Figshare ORD and their usage.

Our analysis highlighted two main findings. First of all, we observe a consistent pattern of under-usage, which surely makes room to wonder why some some ORD records get more attention than others do, and in some cases thousands of ORD records are completely 'invisible' to users. As simple as this finding is, the huge coverage of this sample is throwing evidence over a widespread behavior across formal, natural and social sciences; as well as engineering, technology, design sciences; and last but not least, humanities and languages. The researchers of almost all areas of knowledge, seem to behave similarly in publishing more consistently than looking at or downloading ORD, which easily leads to the assumption of a still low re-use (for replication or for creative purposes) of ORD. However, one should consider few exceptions such as *Engineering*, *Information and Computing Sciences*, *Health Sciences*, *Built environment and design*, *Studies in Human society* (see Figure 6) where the situation is paradoxical (less publications than reads/downloads). One interesting result here reported is that smaller and less traditional research

areas seem to show less unbalanced situations, even if this observation could not be confirmed by inferential statistics.

Our second finding, relating to assessing whether the low propensity to use ORD is associated with a low quality of the metadata (measured with our approach), gave a substantial negative result, by contrast with the current literature. This is coherent with the assumptions relating the complex politics of data friction. Our research opens hence to a number of possibilities as relates to further studies, taking into consideration the following questions: Why some ORD are over consulted and other invisible? Which are the attributes of mostly used ORD? Are there other quality parameters that could be explored in order to align quality definitions by the platforms and by the users? How can automated procedures be advanced in order to improve the sensibility of quality parameters? Which other Open Data repositories could be further explored? And many others, connecting digital infrastructures, classification, automatisation of extraction of ORD quality metrics and the users' social behavior. This complex, interdisciplinary package will be required to advance the research on Open Research Data as expression of utmost importance for the Open Science movement.

As for the limitations in this study, one should take into consideration that the direct use observable through the adopted metrics is surely insufficient at the time of exhausting the potential of the data offered by Figshare. While justifying our selection of Figshare as the platform selected for the study, we mentioned that there are other indirect use parameters that might be deemed more meaningful. An interesting parameter could trace how indirect users relate to third-party applications, along with the number of the same applications. However, if the collection of an application's users is demanding, the number of applications that reuse a dataset could be pursued quite easily as for the French OGD portal. This number can help portal managers measure indirect users of their repositories. To this end, tools should be provided to allow application developers to list the datasets they reuse[74]. Furthermore, this information could increase the awareness of users of these applications towards the provenance of the reused data, thus increasing their reliability[75]. The usage analysis on OGD would support ORD portals' strategies in this sense. As previously reported, in the case of OGD from the European Data Portal there are categories that receive much more attention than others.

While motivations for usage of OGD can be diversified, being the users highly different from researchers as primary users of ORD, one could expect that the "mismatch between available data sets and re-used data could be supported by the development of a publication strategy which is better aligned to the needs of end data users"[46].

In the discussion, we reported several motivations coming out from the literature addressing the usage of ORD as complex social activity (data literacy, research cultures, quality definitions, etc.) In this regard, our article highlights the complexities of researching ORD usage and its quality tracing as emergent characteristic of efficient users' engagement with technological infrastructures. The evolution of research needs to advance infrastructures' affordances to put them closer to the users, but also users' need to be endowed to exploit such affordances by qualifying them to engage with advanced notion of usage and quality.

References

1. European Commission. Digital science in Horizon 2020. Technical report, European Commission, Brussels, 2013. URL ec.europa.eu/information{_}society/ newsroom/cf/dae/document.cfm?doc{_}id=2124.

2. Fecher B and Friesike S. Open science: one term, five schools of thought. In *Opening Science*. 2014. pp. 17–47.

3. Veletsianos G and Kimmons R. Scholars in an increasingly open and digital world: How do education professors and students use Twitter? *The Internet and Higher Education* 2016; 30: 1–10. DOI:10.1016/j.iheduc.2016.02.002.

4. Gregory KM, Cousijn H, Groth P et al. Understanding data search as a socio-technical practice. *Journal of Information Science* 2019; 0(0): 0165551519837182. DOI:10.1177/ 0165551519837182.

5. Molloy JC. The open knowledge foundation: Open data means better science. *PLoS Biology* 2011; 9(12). DOI:10.1371/journal.pbio.1001195.

6. Lyon L. Transparency: the emerging third dimension of open science and open data. *The Liber Quarterly* 2016; 25.

7. Borgman CL. *Big Data, Little Data, No data: Scholarship in the Networked World*. Cambridge, MA.: MIT Press, 2015. ISBN 978-0-262-02856-1.

8. Lourenço RP. An analysis of open government portals: A perspective of transparency for accountability. *Government Information Quarterly* 2015; 32(3): 323 – 332. DOI:https: //doi.org/10.1016/j.giq.2015.05.006.

9. Palmirani M, Martoni M and Girardi D. Open government data beyond transparency. In Kő A and Francesconi E (eds.) *Electronic Government and the Information Systems Perspective*. Cham: Springer International Publishing. ISBN 978-3-319-10178-1, pp. 275–291.

10. Màchovà R and Lnenicka M. Evaluating the Quality of Open Data Portals on the National Level. *Journal of theoretical and applied electronic commerce research* 2017; 12: 21 – 41.

11. Wilkinson MD, Dumontier M, Aalbersberg IJ et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 2016; 3: 160018. DOI:10.1038/sdata.2016. 18.

12. Janssen M, Charalabidis Y and Zuiderwijk A. Benefits, adoption barriers and myths of open data and open government. *Information Systems Management* 2012; 29(4): 258–268. DOI: 10.1080/10580530.2012.716740.

13. Science D, Hahnel M, Fane B et al. The state of open data report 2018, 2018. DOI: 10.6084/m9.figshare.7195058.v2.

14. Zuiderwijk A, Janssen M and Susha I. Improving the speed and ease of open data use through metadata, interaction mechanisms, and quality indicators. *Journal of Organizational Computing and Electronic Commerce* 2016; 26(1-2): 116–146. DOI:10.1080/10919392.2015. 1125180.

15. Reiche K and Hofig E. Implementation of metadata quality metrics and application on public government data. pp. 236–241. DOI:10.1109/COMPSACW.2013.32.

16. Neumaier S, Umbrich J and Polleres A. Automated quality assessment of metadata across open data portals. *J Data and Information Quality* 2016; 8(1): 2:1–2:29. DOI:10.1145/2964909.

17. Bates J. The politics of data friction. *Journal of Documentation* 2017; 74. DOI:10.1108/ JD-05-2017-0080.

18. Sadiq S and Indulska M. Open data: Quality over quantity. *International Journal of Information Management* 2017; 37(3): 150–154.

19. Veletsianos G. A Case Study of Scholars' Open and Sharing Practices. *Open Praxis* 2015; 7(3): 199–209. DOI:10.5944/openpraxis.7.3.206.

20. Wouters P and Haak W. Open data : The researcher perspective. Technical report, Leiden, 2017. DOI:10.17632/bwrnfb4bvh.1.

21. Edwards PN, Mayernik MS, Batcheller AL et al. Science friction: data, metadata, and collaboration. *Social studies of science* 2011; 41 5: 667–90.

22. Anonymous1. Details omitted for double-blind reviewing.

23. Anonymous2. Details omitted for double-blind reviewing.

24. European Commission - H2020 Programme. Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020. Technical report, European Commission, Brussels, 2017. URL http://ec.europa.eu/research/ participants/data/ref/h2020/grants{_}manual/hi/oa{_}pilot/ h2020-hi-oa-pilot-guide{_}en.pdf.

25. Wellcome Trust. Wellcome signs open data concordat. *Wellcome Trust Blog* 2016; URL https://wellcome.ac.uk/news/ wellcome-signs-open-data-concordat.

26. NWO. Open Science. URL https://www.nwo.nl/en/policies/open+science.

27. CERN. CMS data preservation, re-use and open access policy. CERN Open Data Portal, 2018. DOI:10.7483/OPENDATA.CMS.7347.JDWH. URL http://opendata.cern. ch/record/414.

28. Bill & Melinda Gates Foundation. Gates Open Research, 2017. URL https:// gatesopenresearch.org/about/policies{#}dataavail.

29. European Commission - RISE - Research Innovation and Science Policy Experts. Mallorca declaration on open science: Achieving open science, 2016. URL https://ec.europa.

eu/research/openvision/pdf/rise/mallorca{_}declaration{_}2017.
pdf.

30. Minniti S, Santoro V and Belli S. Mapping the development of Open Access in Latin America and Caribbean countries. An analysis of Web of Science Core Collection and SciELO Citation Index (2005–2017). *Scientometrics* 2018; 117(3): 1905–1930. DOI: 10.1007/s11192-018-2950-0.

31. McKiernan EC, Bourne PE, Brown CT et al. How open science helps researchers succeed. *eLife* 2016; 5. DOI:10.7554/eLife.16800. URL http://www.ncbi.nlm.nih.gov/pubmed/27387362.

32. Lämmerhirt D. Briefing Paper : Disciplinary differences in opening research data. *Pasteur4Oa* 2016; (June 2013): 1–8.

33. Zuiderwijk A and Janssen M. Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly* 2014; 31(1): 17–29. DOI: 10.1016/j.giq.2013.04.003.

34. Dai Q, Shin E and Smith C. Open and inclusive collaboration in science: A framework. *OECD Science, Technology and Industry Working Papers* 2018; 2018(7). DOI:doi.org/10.1787/2dbff737-en.

35. Gurstein MB. Open data: Empowering the empowered or effective data use for everyone? *First Monday* 2011; 16(2): 1–8. DOI:10.1177/0170840601223003.

36. Zuiderwijk A, Janssen M and Dwivedi YK. Acceptance and use predictors of open data technologies: Drawing upon the unified theory of acceptance and use of technology. *Government Information Quarterly* 2015; 32(4): 429–440. DOI:10.1016/j.giq.2015.09.005.

37. Schneider R. Research Data Literacy. In *Communications in Computer and Information Science*, volume 397 CCIS. Springer Verlag. ISBN 9783319039183, pp. 134–140. DOI: 10.1007/978-3-319-03919-0_16.

38. Pouchard L and Bracke MS. An analysis of selected data practices: A case study of the purdue college of agriculture. *Issues in Science and Technology Librarianship* 2016; 2016(85). DOI: 10.5062/F4057CX4.

39. Wiorogórska Z, Leśniewski J and Rozkosz E. Data Literacy and Research Data Management in Two Top Universities in Poland. Raising Awareness. In *Communications in Computer and Information Science*, volume 810. Springer, Cham. ISBN 9783319743332, pp. 205–214. DOI: 10.1007/978-3-319-74334-9_22.

40. Vilar P and Zabukovec V. Research data management and research data literacy in Slovenian science. *Journal of Documentation* 2019; 75(1): 24–43. DOI:10.1108/JD-03-2018-0042.

41. Carlson J, Fosmire M, Miller C et al. Determining Data Information Literacy Needs: A Study of Students and Research Faculty. *portal: Libraries and the Academy* 2011; 11(2): 629–657. DOI:10.1353/pla.2011.0022.

42. Teal TK, Cranston H Karen Aand Lapp, White E et al. Data Carpentry: Workshops to Increase Data Literacy for Researchers. *International Journal of Digital Curation* 2015; 10(1): 135–143. DOI:10.2218/ijdc.v10i1.351.

43. Hahnel M, Treadway J, Fane B et al. The State of Open Data Report 2017. Technical report, Figshare, London, 2017. DOI:https://doi.org/10.6084/m9.figshare.5481187.v1.

44. Fane B, Ayris P, Hahnel M et al. The State of Open Data Report 2019: a selection of analyses and articles about open data, curated by Figshare. Technical report, 2019. DOI: 10.6084/M9.FIGSHARE.9980783.V2.

45. Open Science Monitor. Facts and Figures for open research data. URL https://ec.europa.eu/info/research-and-innovation/strategy/goals-research-and-innovation-policy/open-science/open-science-monitor/facts-and-figures-open-research-data{_}en.

46. Berends J, Carrara W, Engbers W et al. *Reusing Open Data: a study on companies transforming Open Data into economic and societal value*. European Union, 2017. URL https://www.europeandataportal.eu/sites/default/files/re-using_open_data.pdf.

47. Sasse T, Smith A, Broad E et al. Recommendations for open data portals: from setup to sustainability., 2017. URL https://www.europeandataportal.eu/sites/default/files/edp_s3wp4_sustainability_recommendations.pdf.

48. Amorim RC, Castro JaA, Rocha Da Silva Ja et al. A comparison of research data management platforms: Architecture, flexible metadata and interoperability. *Univers Access Inf Soc* 2017; 16(4): 851–862. DOI:10.1007/s10209-016-0475-y.

49. Ouzzani M, Papotti P and Rahm E. Introduction to the special issue on data quality. *Information Systems* 2013; 38(6): 885–886. DOI:10.1016/j.is.2013.03.001.

50. Batini C and Scannapieco M. *Data and Information Quality - Dimensions, Principles and Techniques*. Data-Centric Systems and Applications, Springer, 2016. ISBN 978-3-319-24104-3. DOI:10.1007/978-3-319-24106-7.

51. Wang RY and Strong DM. Beyond accuracy: What data quality means to data consumers. *J of Management Information Systems* 1996; 12(4): 5–33.

52. Bizer C and Cyganiak R. Quality-driven information filtering using the wiqa policy framework. *Web Semant* 2009; 7(1): 1–10. DOI:10.1016/j.websem.2008.02.005.

53. Batini C, Cappiello C, Francalanci C et al. Methodologies for data quality assessment and improvement. *ACM Comput Surv* 2009; 41(3). DOI:10.1145/1541880.1541883.

54. Kim KS and Sin SCJ. Selecting quality sources: Bridging the gap between the perception and use of information sources. *Journal of Information Science* 2011; 37(2): 178–188. DOI: 10.1177/0165551511400958.

55. Quarati A, Albertoni R and De Martino M. Overall quality assessment of skos thesauri: An AHP-based approach. *Journal of Information Science* 2017; 43(6): 816–834. DOI: 10.1177/0165551516671079.

56. Zaveri A, Rula A, Maurino A et al. Quality assessment for linked data: A survey. *Semantic Web* 2016; 7(1): 63–93. DOI:10.3233/SW-150175.

57. Oliveira MIS, de Oliveira HR, Oliveira LA et al. Open government data portals analysis: The brazilian case. In *Proceedings of the 17th International Digital Government Research Conference on Digital Government Research*. dg.o '16, New York, NY, USA: ACM. ISBN 978-1-4503-4339-8, pp. 415–424. DOI:10.1145/2912160.2912163.

58. Vetró A, Canova L, Torchiano M et al. Open data quality measurement framework: Definition and application to open government data. *Government Information Quarterly* 2016; 33(2): 325 – 337. DOI:https://doi.org/10.1016/j.giq.2016.02.001.

59. Zhu X and Freeman MA. An evaluation of U.S. municipal open data portals: A user interaction framework. *Journal of the Association for Information Science and Technology* 2019; 70(1): 27–37. DOI:10.1002/asi.24081.

60. Wilkinson MD, Sansone SA, Schultes E et al. A design framework and exemplar metrics for fairness. *bioRxiv* 2017; DOI:10.1101/225490.

61. Ubaldi B. Open government data 2013; (22). DOI:https://doi.org/https://doi.org/10.1787/5k46bj4f03s7-en.

62. Boudreau C. Reuse of open data in quebec: from economic development to government transparency. *International Review of Administrative Sciences* 2020; 0(0): 0020852319884628. DOI:10.1177/0020852319884628.

63. Leydesdorff L and Milojević S. Scientometrics. *International Encyclopedia of the Social & Behavioral Sciences: Second Edition* 2012; : 322–327 1208.4566.

64. Ngai EW, Tao SS and Moon KK. Social media research: Theories, constructs, and conceptual frameworks. *International Journal of Information Management* 2015; 35(1): 33–44. DOI: 10.1016/j.ijinfomgt.2014.09.004.

65. Konkiel S and Scherer D. New opportunities for repositories in the age of altmetrics. *Bull Assoc Inf Sci Technol* 2013; 39(4): 22–26. DOI:10.1002/bult.2013.1720390408.

66. Safarov I, Meijer AJ and Grimmelikhuijsen S. Utilization of open government data: A systematic literature review of types, conditions, effects and users. *Information Polity* 2017; 22: 1–24.

67. Khalilzadeh J and Tasci AD. Large sample size, significance level, and the effect size: Solutions to perils of using big data for academic research. *Tourism Management* 2017; 62: 89 – 96. DOI:https://doi.org/10.1016/j.tourman.2017.03.026.

68. Zhang L and Watson E. The prevalence of green and grey open access: Where do physical science researchers archive their publications? *Scientometrics* 2018; DOI:10.1007/s11192-018-2924-2.

69. Lee J, Oh S, Dong H et al. Motivations for self-archiving on an academic social networking site: A study on researchgate. *Journal of the Association for Information Science and Technology* 2019; 70(6): 563–574. DOI:10.1002/asi.24138.

70. Stone A. Are open data efforts working? *government technology* 2018; URL http://www.govtech.com/data/Are-Open-Data-Efforts-Working.html.

71. Merton RK. The normative structure of science. In Merton RK (ed.) *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago, Illinois: University Chicago Press. ISBN 0-226-52092-7., 1973.

72. Lave J and Wenger E. *Situated learning: Legitimate peripheral participation*, *Learning in doing*, volume 95. Cambridge University Press, 1991. ISBN 0521423740. DOI: 10.2307/2804509.

73. Harvey L and Green D. Defining Quality. *Assessment & Evaluation in Higher Education* 1993; 18(1): 9–34. DOI:10.1080/0260293930180102.

74. Vander Sande M, Portier M, Mannens E et al. Challenges for open data usage: open derivatives and licensing. In *Workshop on Using Open Data, Proceedings*. p. 4.

75. Anonymous1. Details omitted for double-blind reviewing.