Data Retrieval = Text Retrieval?

Maryam Bugaje[1] and Gobinda Chowdhury[1]

[1] iSchool, Faculty of Engineering and Environment, Northumbria University, Newcastle, UK
{maryam.bugaje, gobinda.chowdhury}@northumbria.ac.uk

Abstract. Due to the comparatively more recent emergence of data retrieval systems than text-based search engines, the former have still yet to attain similar maturity in terms of standards and techniques. Most of the existing solutions for data retrieval are more or less makeshift adaptations of text retrieval systems rather than purpose-built solutions specially designed to cater to the particular peculiarities, subtleties, and unique requirements of research datasets. In this paper we probe into the key differences between text and data retrieval that bear practical relevance to the retrieval question; these differences we demonstrate by evaluating some representative examples of research data repositories as well as presenting findings from previous studies.

Keywords: Data Retrieval, Text Retrieval, Research Data Management, Research Data Repositories.

1 Introduction

Among the more comprehensive definitions of research data is that they are "entities used as evidence of phenomena for the purposes of research or scholarship", which may range in form from digital records (e.g. text, audio, video, spreadsheets, etc.) to physical objects (e.g. laboratory specimens, historical artefacts, soil samples, etc.) [1]. A stricter definition, however, stipulates that in addition, research data must be associated with useful metadata, or "information describing its creation, transformation, and/or usage context" [2]. Research data repositories perform various useful functions, among the first of which is storage/curation of research datasets, and not the least of which is enabling the discoverability of the same by authorized parties. The latter function is primarily fulfilled by the retrieval system via (a) a search interface by means of which the underlying database may be queried; (b) a browsing interface through which the same may be accomplished in a structured way; or (c) a URL that links directly to the resource itself. Data retrieval systems are still at a relatively early stage of development, and most of the data repositories currently in use are essentially text-based in their methods of metadata indexing, query processing, and retrieval; and also in the way that their search results are presented. Superficially, this fact may hardly be regarded as constituting a definite issue in itself, until the question is considered whether we interact differently with data than with publications; and, if so, whether there may not be better advantage, then, in modelling data retrieval systems specially to reflect the unique requirements and opportunities indicated by these differences. This is an important

question retrieval-wise, partly because the task of tagging research datasets with metadata, which is the central component that powers the retrieval engine, is often complex; and partly because unlike the indexing of research papers by services like Web of Science, the indexing of research datasets is not standardized or controlled [3]. This paper recognizes the need to not only identify existing problem areas in data retrieval, such as the aforementioned; but as well the relationships of these problems to one another, in order that they may be traced to, and addressed at the root. There is need, also, to ascertain the requirements of a proper data retrieval system in order that appropriate means may be devised for the achievement of that end. It is not our object in this paper to expound on the theoretical differences between text retrieval and data retrieval, but rather, to investigate the more evident and frequently encountered differences that bear practical relevance to retrieval. The particular aims of this paper, which form a part of an ongoing research, have been tailored expressly with this purpose in mind; they are:

To –

1. Review the currently supported features and functionalities of RDM repositories as pertains retrieval. It is not part of our aim to critique these services from a usability perspective, or to compare their general features, but to provide a snapshot of the standard search and retrieval features available;
2. Assess the degree to which these services cater and are adapted to the special requirements of data retrieval as distinguished from text retrieval (i.e. research publications);
3. Ascertain as to the existence of any marked improvements in retrieval performance and output, between services that support only simple-keyword searches and those that support more advanced querying options; and
4. Establish, via an exploratory study, the differences, as specially pertains retrieval, between the requirements of research data content and text content.

This paper addresses points 1 and 4 above.

2 Appraisal of Repositories in Current Use

As of date re3data.org lists upwards of 900 research data repositories in its directory. For better manageability of this mammoth number, and for the purpose of giving structure to our review we have organized these into 6 broad, non-mutually exclusive groups viz. disciplinary, institutional, publisher-service, location-based, dedicated content-type, and commercial/general purpose repositories. For each group, we have hand-picked a few representative examples for evaluation against the following yardsticks which have special relevance to the retrieval question:

1. *Metadata.* The method by which the metadata associated with each dataset is extracted and used for indexing; and the degree to which this metadata appears to be exploited to provide features for browsing, searching/querying, filtering and result presentation;

2. *Querying facility*. The level of expressiveness allowed in searching/querying the repository; which, in addition, further enhances discoverability; and
3. *Result filtering.* The availability of options for filtering down search results, and the furthest granularity to which this is possible.

Our choices of data repository examples were guided by the combined recommendations of re3data.org and Nature[1].

Disciplinary Repositories. These are dedicated repositories housing research data from a specific disciplinary branch or sub-branch; e.g. Dryad[2] for the Biosciences, and the Virtual Solar Observatory (VSO)[3] for Solar Physics data. Fig. 1. shows the search interface of the VSO, where metadata can evidently be seen to be made ample use of to enable searches by an extensive range of variables. To be sure, solar data is a highly standardized, machine-collected data; and it is exhaustively machine-tagged to a metadata schema standard to the discipline. The latter affords immense potential for building, on the strength of it, functionalities capable of supporting very expressive search queries, as well as result filtering to a fine granularity. It also allows for better indexing methods, and, consequently, more efficient retrieval.

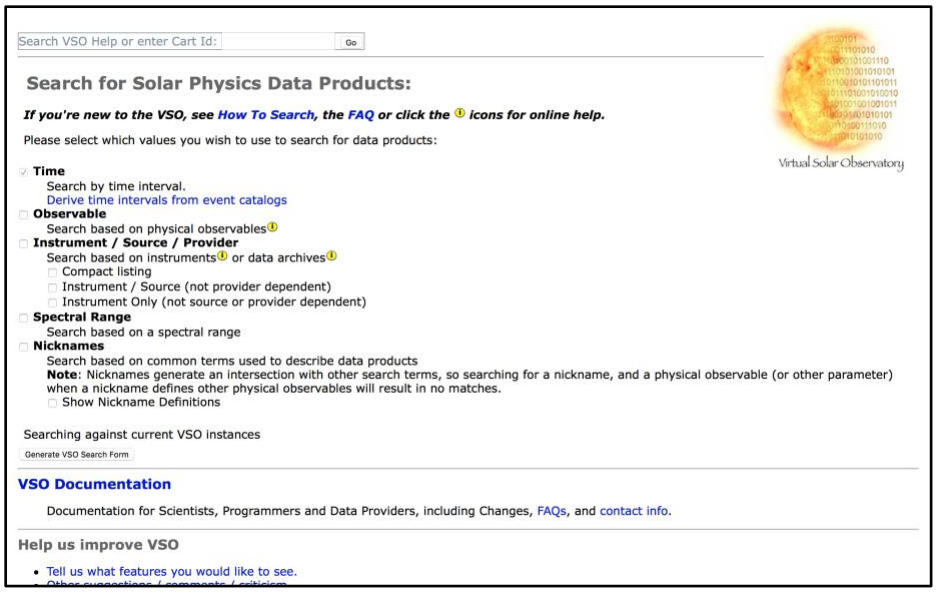

Fig. 1. The bounded domain of disciplinary repositories affords scope for exploiting disciplinary metadata to improve query expressiveness, indexing techniques, and retrieval efficiency among others

1 https://www.nature.com/sdata/policies/repositories
2 datadryad.org/
3 https://sdac.virtualsolar.org/cgi/

Publisher-service Repositories. These are provided by journal publishers, some of whom conduct peer reviews on research data and publish them as regular scholarly outputs; e.g. Nature's Scientific Data[4]. Publisher-service repositories are mostly optimized for linking research data with the publications that they underlie; and, as journals generally publish around specific subjects/topics, their repositories may share some of the aforementioned advantages of disciplinary repositories; these services, however, are few.

Institutional Repositories. Institutions of higher learning may make available repositories for the exclusive use of their research communities; e.g. Oxford University's Research Data Oxford[5]. These repositories are usually hidden behind a login, and many universities outsource the provision of this service to third-party vendors. Furthermore, the repositories are built such that they could as well house other research outputs, including books, patents, reports, and theses among others. All these combine to ultimately give very little scope for specially adapting their retrieval systems to work well for research datasets. As could be seen in Fig. 2., however, institutional repositories may have a modest provision of options for advanced searching and for filtering search results.

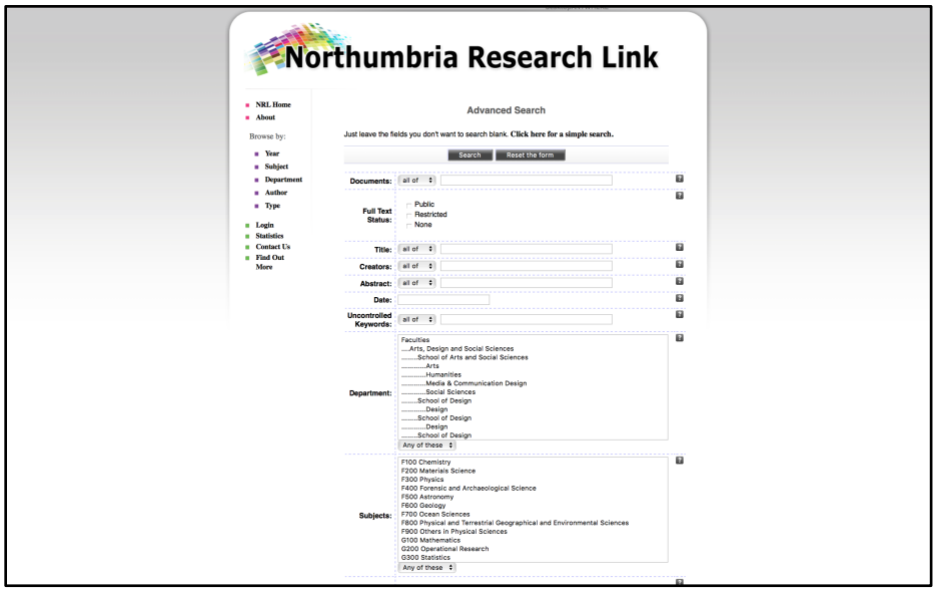

Fig. 2. Institutional repositories are designed, generally, to accommodate other research outputs in addition to datasets. Consequently there is little scope for data-centric features.

Location-based Repositories. Research Data housed in these repositories are generally accessible to anyone globally, but submissions are solicited and accepted only

4 https://www.nature.com/sdata/
5 http://researchdata.ox.ac.uk

from researchers within a specified geographical area; e.g. ANDS Research Data Australia, and the European Union Open Data Portal (EU ODP)[6]. These repositories are generally more data-centric than institutional repositories, and feature advanced search options that are more pertinent to research data (e.g. Fig. 3.); but, in their attempt to accommodate all data that falls within their geographical boundaries, they sacrifice much of the benefits, such as have been previously mentioned under the example of disciplinary repositories, of well-exploited metadata which come with having a more streamlined content.

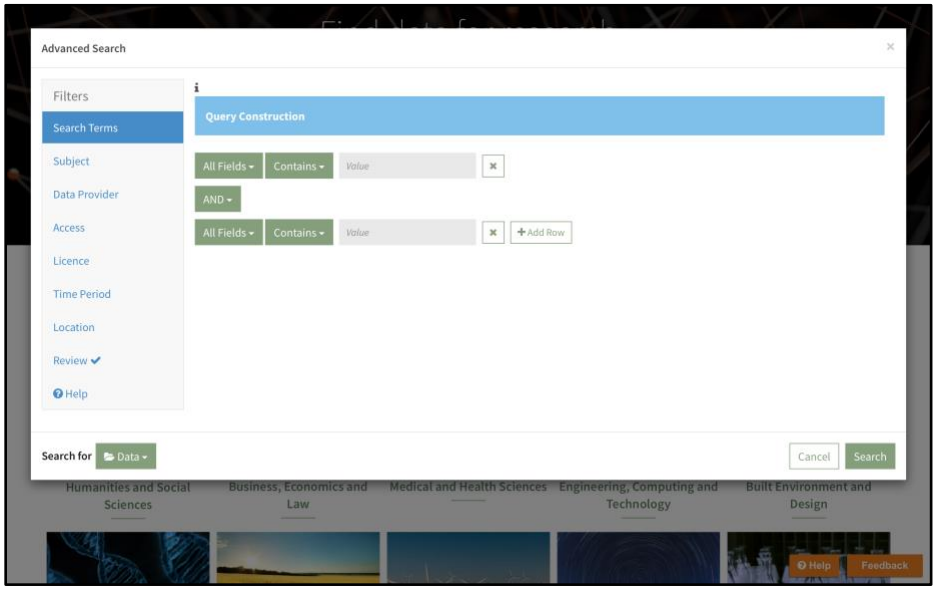

Fig. 3. Advanced search options by Research Data Australia

Commercial Service-provider & General-purpose Repositories. These place little to no restrictions on research data submitted to them; e.g. Figshare[7]. They tend to house multidisciplinary data, as well as data from niche disciplines that do not have dedicated repositories. As shown in Fig. 4., general-purpose repositories, by the mere fact of their being general-purpose, find it harder to achieve any fine-grained filtering of search results, to say nothing of forming expressive queries. This is because the metadata that is needed to support such functionalities is, in the interest of inclusivity, kept superficial at best; and, as such, the retrieval mechanism is essentially very text-like.

Dedicated Content-type Repositories. These exclusively or predominantly house research data of a certain file type/format; e.g. the Visual Arts Data Service (VASD)[8] for image data, shown in Fig. 5. By virtue of the comparative homogeneity of their

6 https://data.europa.eu/euodp/en/data
7 https://figshare.com
8 https://vads.ac.uk

supported data, not only do dedicated content-type repositories have the unique advantage of potentially having their retrieval engines designed specially to cope with their content type and support interaction possibilities unique to it, but also their search interfaces to be designed around ideas and principles as best suit or express the special properties of their content.

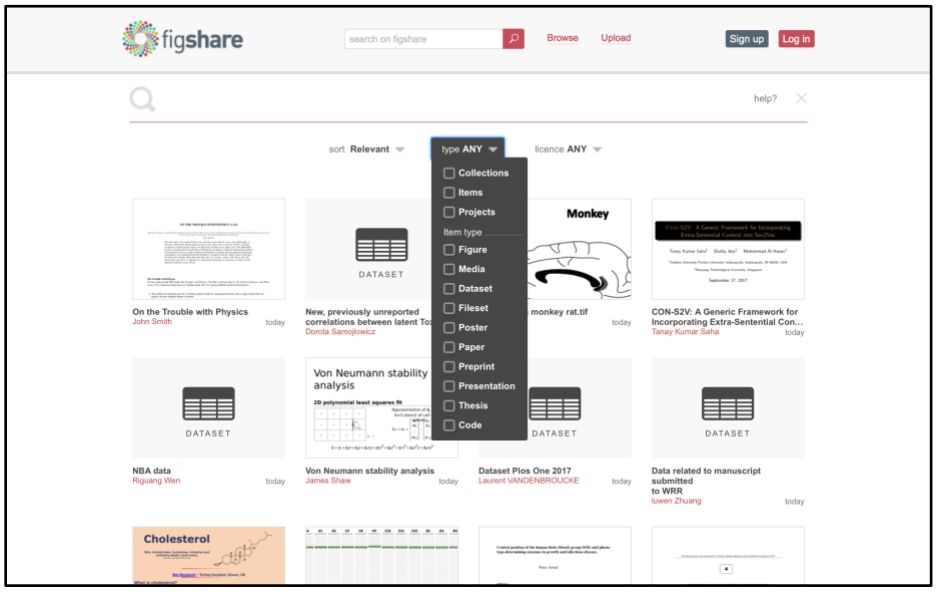

Fig. 4. Showing Figshare as an example of general-purpose/commercial data repositories.

2.1 Section Summary

While in the preceding sections we have dwelt on the strengths and advantages that purpose-built data retrieval systems promise, we have not sufficiently touched on the disadvantages and consequences of settling for a text-based system for data retrieval. Unlike research publications (text), data may be said to entail an active interaction: researchers do not "read" datasets in the passive sense that they do publications; rather, they "use" it by visualizing, combining, or manipulating it among other things. In the section that follows we briefly present the findings of a previous exploratory study that argues a strong case in favor of retrieval solutions designed purposely for use with data [3].

3 Comparison-in-Action Between Text & Data Retrieval

Fig. 6. shows the search interface of the popular Web of Science[9], a text-based search engine for research publications. It can be observed that the search options it provides

9 https://apps.webofknowledge.com/

do not decidedly differ from those previously seen of data repositories. In fact, the resemblance is not superficial: at their core the vast majority of currently-available data-repository retrieval engines are effectively identical to text retrieval engines. This state of affairs is far from ideal because the differences in file types, size, and format, as well as the need for documentation for research datasets, have major implications for search efficiency and resource requirements of data retrieval.

Fundamentally, the basic building block of all scientific publications is text. This uniformity makes it easy to develop standards and fine-tuned solutions. Data, however, even by its mere definition indicates variability. The sheer variation in file types and formats of datasets makes any standardization unfeasible. One of the key challenges of data retrieval arises from the lack of use of standard metadata and documentation to contextualize data sufficiently for re-use [4] and discovery [5,6].

Also, the file sizes of research datasets typically exceed the file sizes of research publications (text). It could be observed from Table 1 that the average file size of a single research dataset may in some disciplines amount to as much as 900 times over the average file size of a single research publication. The ordinary web browser, consequently, cannot support the preview of datasets online as it does research publications; and consequently in turn, datasets must necessarily be downloaded before even a glimpse of them could be had [3]. These false downloads of large files result in considerable processing overhead, and it is more advisable that the retrieval system returns a manageable subset of the data so that the user may view it beforehand and be able make an informed decision as to whether to download it.

Research shows that energy consumption increases with increase in server load because energy is consumed during both phases: while doing computing work and while waiting for database data to arrive [7]. Hence, a reduction in the volume of data downloaded will reduce the energy consumption of IT infrastructure of data services as well as the universities and research institutions, thereby reducing the environmental costs of research data management.

Fig. 5. Showing VADS as an example of dedicated content-type data repositories.

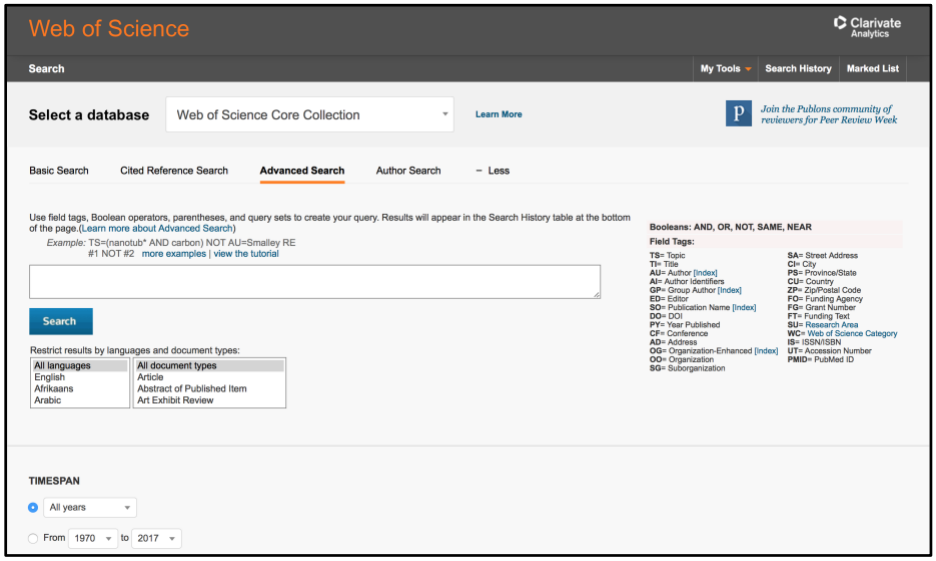

Fig. 6. An example of a text-based retrieval system, Web of Science, showing advanced search options.

Table 1. Average sizes of files retrieved for research datasets and research publications.

Discipline	Keywords	Data Retrieval*	Text Retrieval*	Approx. ratio of text to data
Arts & Humanities	art museums	6.205 MB	0.820 MB	1:8
	nineteenth century	2.898 MB	1.042 MB	1:3
	“world war”	6.158 MB	0.508 MB	1:12
	medieval	5.158 MB	1.091 MB	1:5
	popular music	9.334 MB	1.000 MB	1:9
Social Sciences	unemployment	4.729 MB	0.455 MB	1:10
	cognition	13.340 MB	1.612 MB	1:8
	“labour law”	2.827 MB	0.410 MB	1:7
	“trade union”	15.939 MB	0.748 MB	1:21
	imprisonment	2.444 MB	0.503 MB	1:5
Computer & Information Science	search behavior	657.707 MB	0.731 MB	1:900
	face recognition	**1.394 GB**	1.535 MB	1:908
	computer vision	**1.339 GB**	2.782 MB	1:481
	research data sharing	1.574 MB	0.521 MB	1:3
	social media data	19.597 MB	1.078 MB	1:18
Natural Sciences	marine life	32.318 MB	1.491 MB	1:22
	“climate change”	2.808 MB	2.497 MB	1:1
	“renewable energy”	766.432 MB	3.606 MB	1:213
	“ultraviolet light”	496.745 MB	1.991 MB	1:250

	"oxidative phosphorlyation"	41.177 MB	1.895 MB	1:22

**Average File Size, inclusive of documentation*
***Average File Size*

4 Conclusion

With special reference to retrieval, this paper has expounded on some key differences between research data and publications (text), and urged the development of data retrieval systems that are modelled around requirements and opportunities unique to data. The current state of affairs in which data retrieval and text retrieval are equated and dealt with interchangeably is unsatisfactory, unsustainable (for details on sustainability of information see [8,9]) and results in an unnecessarily high consumption of network, computing, and storage resources. Most of the current data retrieval systems offer features that are based on keyword searches and are appropriate for text retrieval, but they seldom meet the specific requirements of data retrieval;

References

1. Borgman, C. (2015). Big Data, Little Data, No Data: Scholarship in the networked world. The MIT Press.
2. Weber, A.; Piesche, C. Requirements on Long-Term Accessibility and Preservation of Research Results with Particular Regard to Their Provenance. ISPRS Int. J. Geo-Inf. 2016, 5, 49.
3. Bugaje M., Chowdhury G. (2017) Is Data Retrieval Different from Text Retrieval? An Exploratory Study. In: Choemprayong S., Crestani F., Cunningham S. (eds) Digital Libraries: Data, Information, and Knowledge for Digital Lives. ICADL 2017. Lecture Notes in Computer Science, vol 10647. Springer, Cham
4. Borgman, C.L.: The conundrum of sharing research data. Journal of the American Society for Information Science and Technology, 63(6), 1059–1078 (2012).
5. Borgman, C.L., Wallis, J.C., Mayernik, M.S.: Who's got the data? Interdependencies in science and technology collaborations. Computer Supported Cooperative Work 21(6), 485-523 (2012).
6. The Data harvest: How sharing research data can yield knowledge, jobs and growth. An RDA Europe report (December 2014). https://rd-alliance.org/sites/default/files/attachment/The%20Data%20Harvest%20Final.pdf, last accessed 2017/06/11.
7. Boru, D., Kliazovich, D., Granelli, F., Bouvry, P., Zomaya, A.Y.: Energy-efficient data replication in cloud computing datacenters. Cluster Computing, 18(1), 385-402 (2015)
8. Chowdhury, G.G.: Sustainability of scholarly information. Facet Publishing, London, (2014).
9. Chowdhury, G.G.: How to improve the sustainability of digital libraries and information services? Journal of the Association for Information Science and Technology, 67(10), 2379-91 (2016).