
Proceedings Paper:
Koya, K. and Chowdhury, G. (2022) A quality and popularity based ranking method for research datasets. In: APIT 2022: 2022 4th Asia Pacific Information Technology Conference. APIT 2022: 4th Asia Pacific Information Technology Conference, 14-16 Jan 2022, Virtual conference. ACM Conference Proceedings. ACM Digital Library, pp. 103-110. ISBN: 9781450395571.

https://doi.org/10.1145/3512353.3512368

A quality and popularity based ranking method of research datasets

DR KUSHWANTH KOYA

iSchool, Department of Finance, Accounting and Business Systems, College of Business, Technology and Engineering, Sheffield Hallam University.

k.koya@shu.ac.uk

PROF GOBINDA CHOWDHURY

iSchool, Department of Computer and Information Sciences, Faculty of Science, University of Strathclyde.

Gobinda.chowdhury@strath.ac.uk

Research outputs are the final products in the scientific research process and their quality is progressively being evaluated by various methods such as altmetrics, bibliometrics, impact factors and citation count etc. However, a significant component of scientific research involves creating/collecting/curating research datasets and globally, funding agencies and governments are mandating an open access policy on research datasets. Though repositories exist to store the datasets, there is no metricised guidance, indicating the quality of datasets for researchers wishing to reuse. We propose a novel method for ranking and visualising research datasets based on their quality and popularity, constructed through a normalised citation count since the year of origin, total cites and the impact factor of the journals which publish the articles citing the dataset. Additionally, we present the process flow for a proposed digital information system for the access of datasets according to their discipline and rank based on the variables. The proposed method is expected to assist researchers, globally, to choose the right datasets for their research, encourage researchers to share their datasets and promote interdisciplinary research.

CCS CONCEPTS • General and reference • Cross-computing tools and techniques • Evaluation

Additional Keywords and Phrases: Research data, research data management, research data quality

1 INTRODUCTION

1.1 Attribution to Datasets

Research outputs in the form of journal articles, conference papers, books, and performances such as concerts and theatre are the final products of research in various higher education and research institutions (HEIs & RIs). Nowadays, in addition to peer-review, bibliometric indicators of research outputs are increasingly becoming an important part of assessment of the quality of research performed at various HEIs and RIs, thus becoming informants to agencies offering performance-based research funding [1-4]. National research assessment exercises e.g. the United Kingdom's Research Excellence Framework (REF) and Excellence in Research for Australia (ERA) etc., use research outputs as a measure of quality of research, resulting in disbursal of significant amount of funding. Therefore, bibliometrics play a major role in determining the future of scientific research as they influence ranking of HEIs and RIs around the world and research assessment exercises [5. 6]. However, a major part of the research process involves creating/collecting/curating research datasets, which ultimately lead to the results using appropriate methodologies. Relative to the research project, creating a dataset can take from a few days to many years, signifying a substantial economic value, effort and time attached through the research staff. The dataset is also a result of the research, which when made available becomes an asset to researchers, especially secondary researchers who re-use the data for their own research purposes and general dataset peer-review [7-10].

Currently, governments and funding agencies are mandating an open access policy for datasets which result from research they support, and there is an increasing trend to develop a standard practice of citing datasets. Provided a dataset is assigned an identifier i.e. Unique Dataset Identifier (UDI), it becomes relatively simple to track research outputs citing the dataset, thereby allowing the evaluation of impact of a particular dataset. Many products have resulted through this perspective i.e. Nature Publication Group's Scientific Data, Datacite.org, Inter-university Consortium for Political and Social Research (ICPSR) and Thomson Reuters Data Citation Index [11-13]. Additionally, the current available products only measure how many times a certain dataset has been cited by researchers, but do not measure the quality of the dataset in terms of peer-review. The products function with a general presumption that a highly cited dataset is a good quality dataset. Another solution proposed by researchers is to cite the research output, which is the product of a dataset in place of citing the dataset. This ensures that the dataset and the research output have been peer-reviewed which assures the reliability of the data. However, this does not inform us the impact of the data. This is the current problem in the measurement of the real impact of data [14]. To resolve the data impact conundrum, we propose a novel method for ranking and visualising research datasets based on their quality and popularity, constructed through a normalised citation count since the year of origin, total cites and the impact factor of the journals which publish the articles citing the dataset. Considering the impact factor of journals where the published articles cite the dataset will assure quality, which is currently missing in products that are currently available to measure the quality of research datasets.

1.2 Encouraging researchers to share datasets

Researchers are increasingly encouraged to share their data in repositories or required to make their datasets publicly available for the purpose of reuse through various initiatives as mentioned earlier. However, the submission rates are low as researchers don't feel valued in creating/collecting/curating research data-sets, in addition to the lack of standards and workflows, and the consequences of improper use [10, 15-17]. Observed hindrances also include lack of citation standards of datasets, discovery and access issues [18]. Organisations i.e. Datacite and re3data have offered robustly constructed repository services and DOIs for research datasets to encourage researchers to share and curate their research data. However, lack of peer-review, citation index, recognition and discoverability issues repel researchers from sharing their data. Hence, it becomes essential to provide a system for researchers where their time and effort are appreciated in the creation and management of datasets. Creating an index where a researcher's dataset is rated based on quality (considering the impact factor of articles using the dataset) and popularity (citations received by the dataset) will encourage researchers to share their datasets.

1.3 Data quality dimensions

Studies on the quality of research datasets is not something new in academia. A comprehensive review reveals the dimensions of data quality (Table 1), which arguably is still valid today shows various indicators of data quality arranged according to their rank based on the number of other researchers who endorsed them [19]. The dimensions are classified into internal view dimensions, concerning design and operation of data, and external view dimensions, concerning use and the value of data.

Table 1. Data quality dimensions (Wand & Wang, 1996)

Dimension	# of citing researchers	Dimension	# of citing researchers	Dimension	# of citing researchers
Accuracy	25	Format	4	Comparability	2
Reliability	22	Interpretability	4	Conciseness	2
Timeliness	19	Content	3	Free of bias	2
Relevance	16	Efficiency	3	Informativeness	2
Completeness	15	Importance	3	Level of detail	2
Currency	9	Sufficiency	3	Quantitativeness	2
Consistency	8	Usableness	3	Scope	2
Flexibility	5	Usefulness	3	Understandability	2
Precision	5	Clarity	2		

Table 2 examines the classification of these dimensions. Internal view examines the design and operation of the data, much like peer-review, and external view examines the use and value of data, much like citation counts.

Table 2. Classification of data quality dimensions (Wand & Yang, 1996)

View	Dimensions
Internal view (design & operation)	Data related – accuracy, reliability, timeliness, completeness, currency, consistency, precision System related – reliability
External view (use & value)	Data related – timeliness, relevance, content, importance, sufficiency, usableness, usefulness, clarity, conciseness, free from bias, informativeness, level of detail, 'quantitativeness', scope, interpretability, understandability System related – timeliness, flexibility, format, efficiency

Similarly, organisations such as W3C and Open Science Framework (OSF) have produced best practices relating to the publication and usage of research data on the Internet. W3C recommends a comprehensive list of 35 factors, which ensure discoverability and comprehensibility of research datasets. The OSF 'badge's datasets for making all components of research openly available and preregistration of all research components at different levels, hence offering recognition to the researchers making their datasets openly available. OSF additionally offers a badge for peer-review of datasets. Thereby, the practices suggested by W3C and OSF encourage researchers to openly share and curate their research datasets. Peer-review is the backbone of scholarly communication of scientific research and is the predominant method in validating research. The peer-review process examines all the data related dimensions and citation counts examine the usage characteristics and popularity of datasets. It is a general perception that journals with a high impact factor (JIF) have high standards of peer-review, in which case, the JIF can be used as a proxy to determine the quality of the dataset until an appropriate measure for research quality is determined by the scientific community. Additionally, there is plenty of literature suggesting the linear relationship between peer-review and journal impact factor [20-24].

Creating a quantifiable rating system taking into account the internal and external views would assist in understanding the quality of a dataset. We thereby propose an index, which indicates the quality of a dataset, considering total dataset citations, average citations since the year of creation and the JIF of the article citing the dataset (Figure 1). For this application to work it is necessary that researchers share their research data along with its UDI (Unique Dataset Identifier) and researchers using the datasets cite the dataset's UDI in their research output to keep track along with the DOI (Digital Object Identifier) and ISBN (International Standard Book Number). The emerging model of e-Science and Open Science encourages sharing of research datasets prior to publication of research outputs to encourage scientific dialogue and collaboration, in addition to measuring quality and popularity [25, 26]. By implementing the proposed index as an engine, it becomes possible to build an application to search (using necessary keywords) and visualise datasets based on quality and popularity.

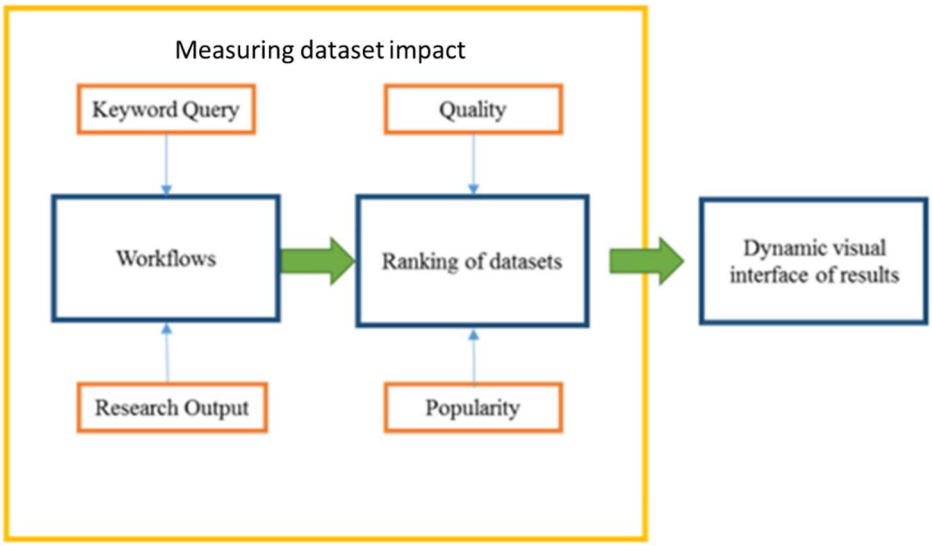

Figure 1: The proposed index

2 METHODOLOGY

In order to build our model, we chose the Physionet.org digital repository as the primary source of various datasets. Physionet's repository consists of various physiological measurements such as EEG, ECG, EHG and gait analysis etc. It is a result of the joint collaboration between Harvard and MIT to facilitate interdisciplinary research between its faculties. Since its inception in the mid-1970s, researchers from around the world have used the datasets for their research purposes [27]. Twenty-three datasets were coded into MS-Excel, including their metadata:

a. Year of publication of the dataset: Helps a potential user to determine the age of the dataset.

b. Total dataset citations: Informs the user regarding the popularity of the dataset.

c. Outputs citing the dataset between the years 2013 and 2015: Gives an overview of popularity in a given period.

d. Year of publication of the citing outputs

e. Publication of the research output and its JIF between the years 2013 and 2015: Establishes the perceived quality of the dataset. The above years were chosen to demonstrate the working of our proposed workflow, although a substantial application considering the entire citation history and JIF can be built. From the JIFs we calculate the Data Impact Factor (DIF), which is the average of all the citing output's publication's JIF in their respective year. Averaging the JIF was the chosen arithmetic as it offers a manageable figure for further analyses.

f. Keywords related to the dataset.

The metadata b to e was obtained from Google Scholar and the JIFs were obtained from Thomson Reuters Journal Citation Reports. Similar data could be collected from any services that provide searchable citation data for research datasets. Keywords were selected by the researchers from the description of the dataset. We calculate the average citations that a dataset received to normalise the citation skewness caused by the year of origin of the dataset. The average citation is calculated by dividing the total citations by the difference between the current year and year of origin of the dataset. Subsequently, we noted the citations count between the years 2013 and 2015 to observe the trend in citations. We further filtered all the citations for journal article submissions and noted the JIFs between the years 2013 and 2015 to observe the trend. For other forms of scholarly communications i.e. books, and conference proceeding etc, there are currently no such numerical indicators of quality, hence, we had to consider it as null. The average of all the JIFs of publications under the datasets citing articles was calculated as a quality indicator. MS Excel was used to calculate the metadata and MS Power BI was used for visualisations of the dataset. As digital object identifiers (DOI) for the citing outputs were irregular and unique dataset identifiers (UDI) are currently inconsistent but evolving, http://www.generatedata.com/ was used to generate random DOIs and UDIs for the citing outputs and datasets.

3 RESULTS

Generally, it was found that the datasets and citations observed in this study were used by researchers in multiple disciplines, either directly using the dataset or citing the corresponding dataset's output to support their study. Table 3 tabulates the various DIF and citation count characteristics of the datasets in alphabetical order.

Here onwards, the index provides an opportunity to researchers regarding their choice of dataset for their research. For example, when one sorts the JIFs for the year 2013 from largest to smallest, it is observed that Dataset 23 possesses the highest DIF in blood pressure data. Dataset 21 also measures blood pressure, however it's DIF for 2013 is 1.89 only, indicating that the quality of Dataset 23 is better than Dataset 21. Additionally, it is also observed that over time (2013 to 2015), the DIFs of both the datasets are increasing, however, Dataset 23 scores are much higher.

In the event of two datasets possessing almost similar DIFs, it is possible to observe the popularity of the datasets by considering the average citations. For example, Dataset 13 and Dataset 16 possess the same DIF for the year 2013. When the average citations for the datasets are observed, Dataset 13 is higher to Dataset 16, also considering that it came into existence in 2006, younger than Dataset 16, which was created in 2003.

4 DISCUSSION

As discussed earlier, this application functions when researchers share their dataset accompanied by its UDI and the researcher using the dataset cites the UDI in their research output. Once developed, other impact measuring indices for books, conference papers and book chapters etc., can be accommodated into the application. Contributing datasets attached with corresponding UDIs into an index will increase the value of research in general through encouraging multidisciplinary research, recognise the authors' effort in creating /collecting /curating the dataset and scholarly communication in general. Once implemented, a secondary researcher in need of a required dataset can either retrieve the dataset from a repository using keywords or retrieve the dataset by following up with the UDI attached to the research output that is being referred.

Table 3. Physionet's datasets and their corresponding metadata

Dataset	Year of Origin	Total Citations	2013 Citations	2014 Citations	2015 Citations	Av Lifetime Citations	DIF 2013	DIF 2014	DIF 2015	Keywords
1. BIDMC Congestive Heart Failure Database	1986	87	4	3	2	2.90	2.10	2.27	2.23	ECG; Congestive heart failure; heart disease; cardiology
2. CEBS Database	2013	6	0	1	4	2.00	1.69	1.78	1.67	ECG; Seismocardiogram; cardiology
3. CHB-MIT Scalp EEG Database	2009	114	26	24	30	16.29	2.06	2.11	2.19	EEG; pediatric EEG; seizure; pediatrics; neurology
4. Congestive Heart Failure RR Interval Database	1995	184	2	7	7	8.76	5.24	5.45	5.61	ECG; Congestive heart failure; heart disease; cardiology; RR interval
5. CAST RR Interval Sub-Study Database	2000	54	0	7	7	3.38	1.85	2.25	2.37	ECG; cardiac arrhythmia; cardiology; RR interval
6. ECG-ID Database	2005	1	0	1	0	0.09	0.00	0.00	0.00	ECG; biometrics
7. EEG Motor Movement/Imagery Dataset	2004	1279	156	169	164	106.58	3.25	3.23	3.19	EEG; motor imagery; neurology; neuroscience; bci2000
8. Effect of Deep Brain Stimulation on Parkinsonian Tremor	2001	33	2	3	2	2.20	2.50	2.68	2.51	EEG; Neuroscience; deep brain stimulation; parkinsons;
9. ERP-based Brain-Computer Interface recordings	2010	24	4	6	5	4.00	2.55	2.43	2.57	EEG; EOG; ERP; bci
10. Evoked Auditory Responses in Normals across Stimulus Level	2010	7	2	2	1	1.17	3.08	3.15	3.39	ABR; OAE; evoked auditory response
11. Exaggerated Heart Rate Oscillations During Two Meditation Techniques	1999	218	25	17	16	12.82	2.52	2.46	2.67	ECG; meditation; athletics; exahherated heart rate
12. Gait Dynamics in Neuro-Degenerative Disease Data Base	1997	552	54	45	45	29.05	2.69	2.74	2.63	Parkinsons; Huntingtons; gait dynamics; EEG
13. Squid Giant Axon Membrane Potential	2006	50	7	1	8	5.00	2.30	2.35	2.37	axons; squid gaint axons; EEG
14. Icelandic 16-electrode Electrohysterogram (EHG) Database	2015	3	0	0	3	3.00	0.00	0.00	0.00	EHG; preterm; pregnancy; gynaecology
15. Noise Enhancement of Sensorimotor Function	2003	375	37	29	47	28.85	2.79	2.78	2.79	posture; motor analysis; gait analysis; sensorimotor
16. Physiologic response to changes in posture	2003	22	4	4	4	1.69	2.30	2.07	2.24	posture; ECG; ABP; gait analysis
17. Post-Ictal Heart Rate Oscillations in Partial Epilepsy: Data and Analysis	1999	53	2	1	1	3.12	1.36	1.36	1.29	ECG; epilepsy; cardiology; neurology; neuroscience
18. The QT Database	1997	294	26	30	31	15.47	1.92	1.92	1.94	ECG; U waves; waveform intervals
19. Santa Fe Time Series Competition Data Set B	1993	48	3	1	0	2.09	1.47	1.39	1.60	multivariate; sleep; polysomnography; heart rate;

										blood oxygen; lung volume
20. Smart Health for Assessing the Risk of Events via ECG	2015	9	0	0	5	9.00	3.19	3.25	4.00	ECG; hypertension; cardiology
21. Stress Recognition in Automobile Drivers	2005	623	91	108	111	56.64	1.89	1.93	1.91	ECG; EMG; galvanic skin resistance; respiration; stress; automobile
22. Term-Preterm EHG Database (TPEHG DB)	2008	59	10	14	20	7.38	1.79	1.78	1.80	EHG; EMG; uterus; pregnancy; premature birth; preterm birth
23. Time Course Data for Blood Pressure in Dahl SS and SSBN13 Rats	2010	24	5	8	4	4.00	8.10	8.55	8.61	rats; hypertension; sodium; baroflex dysfunction; blood pressure

4.1 Retrieval from an index of databases using UDI

To retrieve a required dataset using an UDI, the researcher notes the UDI from any research output which is being referred and searches the dataset index i.e. Physionet.org using the UDI (Figure 2). Once the dataset and its corresponding metadata is retrieved, the researcher has an opportunity to inspect the dataset's quality and popularity parameters by observing the average DIF in a given year and the average citations since the creation of the dataset.

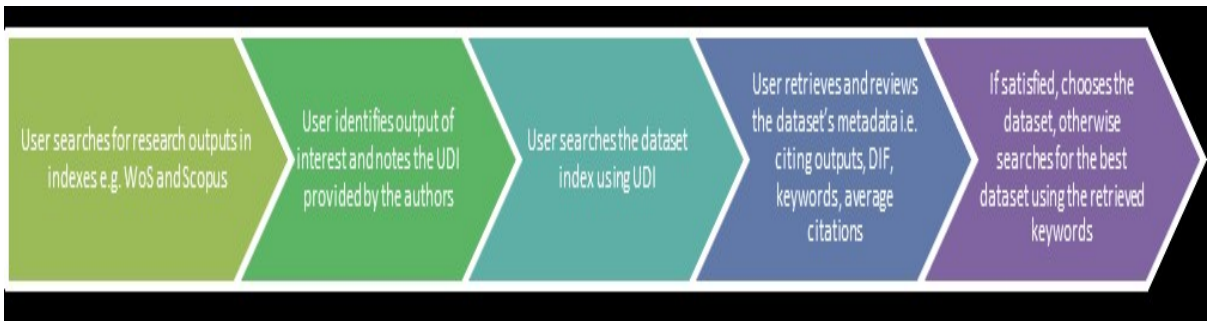

Figure 2: Retrieval of datasets from an index of databases

If satisfied with the quality, the researcher chooses the dataset for their research, if not, the researcher can select the keywords to retrieve datasets under similar classifications to compare the metadata and choose the desired dataset.

4.2 Retrieval from a dataset index using keywords

In the event where the repository hosting datasets is known, the researcher queries using keywords. All datasets attached to the keyword are retrieved along with their corresponding metadata like average citations and average DIF for a particular year, allowing the researcher to choose the datasets which fulfil their requirements (Figure 3).

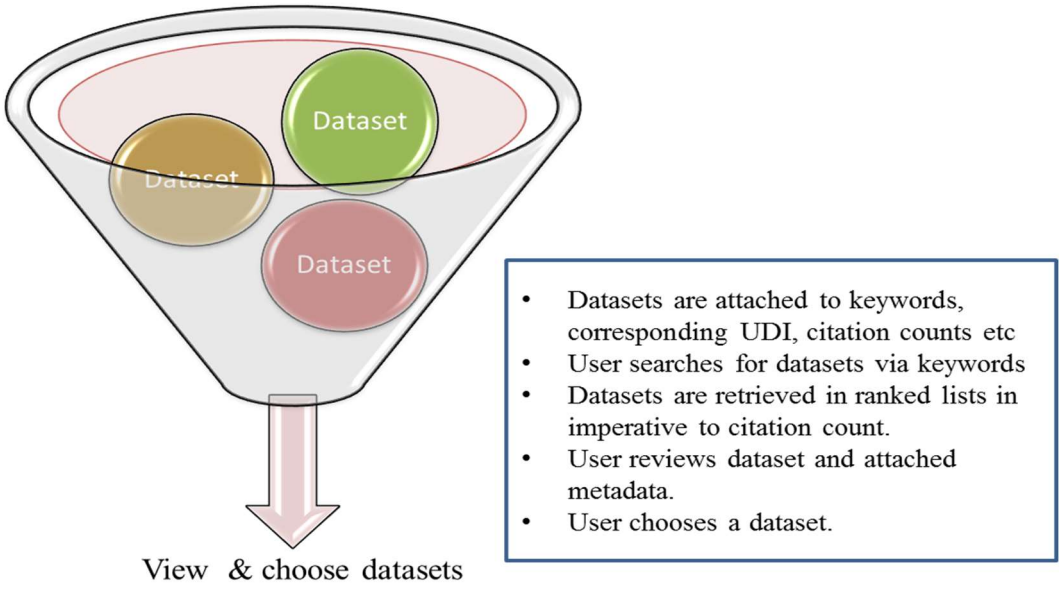

Figure 3: Retrieval of datasets using keywords

Figure 4 observes the visual interface of the results of our proposed application. The keywords i.e. EEG and ECG are connected to the related datasets, with the thickness of the connecting line indicating either the DIF of a dataset in a particular year and the average citations received by the dataset. A dynamic version of Figure 4 can be accessed here. This allows researchers to confidently take a data driven decision in choosing datasets for their research.

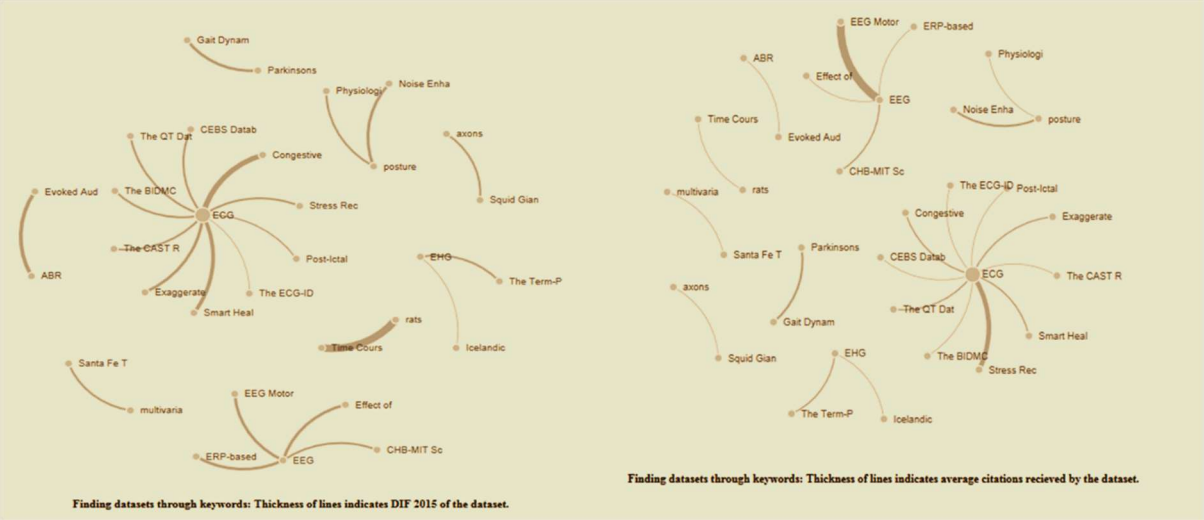

Figure 4: Retrieval of datasets using keywords and filters

4.3 Practical applications of the proposed index

Secondary researchers are the main beneficiaries of the index as it informs them regarding the quality and the popularity of the datasets which they wish to use for their research. Furthermore, the index can contribute towards the benefits of research data sharing as mandated by several research funding governing bodies [28].

4.3.1 Research assessment

Recent developments in research assessment have focused on the impact of research in addition to creation of knowledge [29, 30]. In such circumstances, the proposed index can help measuring the impact of a dataset in the research domain; a dataset can lead to several studies in different disciplines, which demonstrates a significant contribution of the dataset to science in general. Such information can be captured in methods of research assessment i.e. an impact case study of the UK's REF. The ranked datasets are open to further open peer-review, encouraging scientific conversations with regards to dataset quality.

4.3.2 Collaboration

Ranking and identifying reuse of research datasets could result in high level collaboration between different researchers [14]. Visibility of the researchers is improved by increased citations and scientific reputation, in addition to opportunities to discover new methods of capturing and analysing data. It becomes relatively simple to track researchers working in specific areas, encouraging dialogue leading to good practices in data collection methods and knowledge creation being some of the benefits. Additionally, a data management system could be implemented tracking the metadata of the document like infrastructure used to collect data, workflow, ORCIDs of researchers and ISO standards [31, 32]. Re-use of ranked data by researchers in other disciplines promotes interdisciplinary research and a substantial reduction in funds wastage caused by data duplication.

4.3.3 Knowledge economy

The open-access enabled datasets when pooled and attached to metadata, contributes towards creation, evaluation and sharing of datasets considering the various dimensions of data quality, encouraging researchers

to create knowledge in the form of research outputs and linking to datasets utilized. It thus creates a complete knowledge creation circle.

4.3.4 e-Science

The proposed model promotes e-Science by creating a framework for data driven decision making processes, fundamental to grid computing technologies used in organisations around the world e.g. Large Hadron Collider at CERN and Cyberinfrastructure projects of the National Science Foundation [33].

4.4 Comparing the proposed index with existing applications

Table 4. Proposed index compared to existing scientific data operators

Index/ Service	Ranking on popularity (citations)	Ranking on impact factor of users' publications	Visualisation of rankings	Keyword search	Peer-review	No. of citations
Proposed index	✓	✓	✓	✓	✓	✓
Thomson Reuters' Data Citation Index				✓	✓	✓
OSF				✓	✓	
W3C				✓		
Datacite				✓		
re3Data				✓		
Scientific Data				✓	✓	✓

Existing scientific data operators offer various services i.e. peer-review, citation count, discoverability and identifier numbers. However, the proposed index can offer ranking of keyword specific research datasets based on citation count and the reputation of the journals that publish items using the datasets.

5 CONCLUSION

The proposed system can rank datasets by measuring their quality and popularity, which can be implemented into a data intelligence tool to create a visual search platform assisting researchers to select datasets for their research based on quality and popularity. Thus, promoting the sharing, reuse and open discussion of research datasets. This encourages datasets being used beyond its discipline of origin, amalgamation of datasets into a single dataset to create large studies and an open review process through which new data collection and analyses can evolve, thus advancing science in general. The next steps in this research work will involve the creation of a web-based application to implement the proposed index.

FUNDING & CONFLICT OF INTEREST

This work did not receive any form of funding. The authors confirm no conflict of interest.

DATASETS

The datasets are openly available at http://nrl.northumbria.ac.uk/id/eprint/27426

REFERENCES

[1] Liz Allen, Ceri Jones, Kevin Dolby, David Lynn, and Mark Walport. 2009. "Looking for landmarks: the role of expert review and bibliometric analysis in evaluating scientific publication outputs." PlosOne 4, no. 6 : e5910.

[2] Ludo Waltman, Clara Calero-Medina, Joost Kosten, Ed CM Noyons, Robert JW Tijssen, Nees Jan van Eck, Thed N. van Leeuwen, Anthony FJ van Raan, Martijn S. Visser, and Paul Wouters. 2012. "The Leiden Ranking 2011/2012: Data collection, indicators, and interpretation." Journal of the American society for information science and technology 63, no. 12: 2419-2432.

[3] Lutz Bornmann, Rüdiger Mutz, Christoph Neuhaus, and Hans-Dieter Daniel. 2008. "Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results." Ethics in science and environmental politics 8, no. 1: 93-102.

[4] Henk F Moed. 2007. "The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review." Science and Public Policy 34, no. 8: 575-583.

[5] Giovanni Abramo, Ciriaco Andrea D'Angelo, and Flavia Di Costa. 2011. "National research assessment exercises: a comparison of peer review and bibliometrics rankings." Scientometrics 89, no. 3: 929-941.

[6] Antony van Raan. 1996. "Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises." Scientometrics 36, no. 3: 397-420.

[7] Christopher W Belter. 2014. "Measuring the value of research data: A citation analysis of oceanographic data sets." PlosOne 9, no. 3: e92590.

[8] Mark J Costello. 2009. "Motivating online publication of data." BioScience 59, no. 5: 418-427.

[9] Mark A Parsons, Ruth Duerr, and Jean-Bernard Minster. 2010. "Data citation and peer review." Eos, Transactions American Geophysical Union 91, no. 34: 297-298.

[10] Christine L Borgman. 2012. "The conundrum of sharing research data." Journal of the American Society for Information Science and Technology 63, no. 6: 1059-1078.

[11] Scientific Data. 2016. About Scientific Data. Available at www.nature.com/sdata/about

[12] Datacite. About Datacite. 2016. Available at www.datacite.org/about-datacite/what-do-we-do

[13] Thomson Reuters. 2016. Data Citation Index. Available at https://tinyurl.com/yx3657cw

[14] Alex Ball, Monica Duke. 2015. 'How to Track the Impact of Research Data with Metrics'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: https://tinyurl.com/ewcpkvjw

[15] Elisa Bertino. 2014. "Data trustworthiness—approaches and research challenges." In Data privacy management, autonomous spontaneous security, and security assurance, pp. 17-25. Springer, Cham.

[16] RDA Alliance. 2016. Case Statement of the RDA-WDS Publishing Data Interest Group. Available at https://tinyurl.com/48u9crm5

[17] Carol Tenopir, Suzie Allard, Kimberly Douglass, Arsev Umur Aydinoglu, Lei Wu, Eleanor Read, Maribeth Manoff, and Mike Frame. 2011. "Data sharing by scientists: practices and perceptions." PlosOne 6, no. 6: e21101.

[18] Michael Diepenbroek, Hannes Grobe, and Uwe Schindler. 2013. "Research data enters scholarly communication-toward an infrastructure for data publication in the empirical sciences." Thomson Reuters. Available at http://wokinfo.com/products_tools/multidisciplinary/dci/

[19] Yair Wand and Richard Y. Wang. 1996. "Anchoring data quality dimensions in ontological foundations." Communications of the ACM 39, no. 11: 86-95.

[20] Gobinda Chowdhury, Kushwanth Koya, and Pete Philipson. 2016. "Measuring the impact of research: Lessons from the UK's Research Excellence Framework 2014." PlosOne 11, no. 6: e0156978.

[21] Ludo Waltman. 2016. "A review of the literature on citation impact indicators." Journal of informetrics 10, no. 2: 365-391.

[22] Elisabeth S Vieira, José AS Cabral, and José ANF Gomes. 2014. "How good is a model based on bibliometric indicators in predicting the final decisions made by peers?." Journal of Informetrics 8, no. 2: 390-405.

[22] Upali W Jayasinghe, Herbert W. Marsh, and Nigel Bond. 2001. "Peer review in the funding of research in higher education: The Australian experience." Educational Evaluation and Policy Analysis 23, no. 4: 343-364.

[23] Giovanni Abramo, Ciriaco Andrea D'Angelo, and Alessandro Caprasecca. 2009. "Allocative efficiency in public research funding: Can bibliometrics help?." Research policy 38, no. 1: 206-215.

[24] Michael R Nelson. 2009. "Building an open cloud." Science 324, no. 5935: 1656-1657.

[25] Tony Hey, and Anne E. Trefethen. 2005. "Cyberinfrastructure for e-Science." Science 308, no. 5723: 817-821.

[26] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals." circulation 101, no. 23: e215-e220.

[27] EPSRC. 2016. Scope and Benefits of EPSRC policy framework on research data. EPSRC. Available at https://www.epsrc.ac.uk/about/standards/researchdata/scope/

[28] Stephen Merrill and Steve Olson. 2011. Measuring the impacts of federal investments in research: A workshop summary. National Academies Press.

[29] REF. 2014. About the REF. Available at http://www.ref.ac.uk/about/.

[30] ORCID. 2016. What is ORCID? Available at http://orcid.org/about/what-is-orcid/mission .

[31] ISO. 2016. About ISO. Available at http://www.iso.org/iso/home/about.htm.

[32] European Commission. 2016 The European Open Science Cloud: EOSC Infoday, European Commission. Available at https://tinyurl.com/p8hn87cx

[33] Chris Jordan, Maria Esteva, David Walling, Tomilsav Urban, and Sivakumar Kulasekaran. 2013. "Responses to Data Management Requirements at the National Scale." Research Data Management: 64.