A Hybrid Recommendation Approach for Open Research Datasets

Anusuriya Devaraju
CSIRO Mineral Resources
Kensington, Western Australia, 6151
anusuriya.devaraju@csiro.au

Shlomo Berkovsky
CSIRO Data61
Eveleigh, New South Wales, 2015
shlomo.berkovsky@csiro.au

ABSTRACT

Open data initiatives and policies have triggered a dramatic increase in the volume of available research data. This, in turn, has brought to the fore the challenge of helping users to discover relevant datasets. Research data repositories support data search primarily through keyword search and faceted navigation. However, these mechanisms may suit users, who are familiar with the structure and terminology of the repository. This raises the problem of personalized dataset recommendations for users unfamiliar with the repository or not able to clearly articulate their information needs. To this end, we present and evaluate in this paper a recommendation approach applied to a new task — recommending research datasets. Our approach hybridizes content-based similarity with item-to-item co-occurrence, tuned to a feature weighting model obtained through a survey involving real users. We applied the approach in the context of a live research data repository and evaluated it in a user study. The obtained user judgments reveal the ability of the proposed approach to accurately quantify the relevance of datasets and they constitute an important step towards developing a practical dataset recommender.

CCS CONCEPTS

•Information systems → Content ranking; Collaborative filtering; Similarity measures; Recommender systems; •Applied computing → Digital libraries and archives;

KEYWORDS

Recommender system; content-based filtering; item-to-item similarity; open research data; user judgment; digital library.

ACM Reference format:
Anusuriya Devaraju and Shlomo Berkovsky. 2018. A Hybrid Recommendation Approach for Open Research Datasets . In *Proceedings of 26th Conference on User Modeling, Adaptation and Personalization, Singapore, Singapore, July 8–11, 2018 (UMAP '18)*, 5 pages.
DOI: 10.1145/3209219.3209250

1 INTRODUCTION

The Open Science paradigm has led to rapid proliferation of open research data on the Web [18]. Examples of research data include observations, measurements, model outputs, statistics and survey outcomes. Various discipline-specific and common repositories have been established to simplify dissemination of research data. For example, the Registry of Research Data Repository[1] has recorded more than 2000 repositories from various disciplines, and the DataCite portal[2] currently provides access to more than 3.8 millions research datasets.

A number of recent studies [4, 6, 8, 12, 23] have revealed that the current data repositories lack effective data discovery solutions that accurately deliver relevant datasets. Currently, users may find datasets-of-interest in the repositories through keyword search and faceted navigation. These mechanisms may be appropriate for users performing known-item searches, e.g., search by author, title or Digital Object Identifier (DOI), or users, who are familiar with the nature and structure of the repository. However, when users are unable to clearly articulate their needs, seek for datasets in an unfamiliar domain or merely address their ephemeral needs, such search mechanisms may be inadequate [24]. Also, since the search primarily relies on a data description (metadata), top-ranked search results may belong to the same collection and fall short in uncovering novel datasets. These challenges reflect the emergent need for a data discovery solution that complements the search functionality and can deliver personalized dataset recommendations to users.

In this paper, we set out to develop a recommendation approach for open research datasets. Although our approach leverages established recommendation methods, it applies them to a new problem of dataset recommendations. The proposed approach identifies datasets of relevance using a hybrid similarity function that incorporates properties of datasets, e.g., metadata features, as well as their usage patterns, e.g., search/download co-occurrence. The approach is underpinned by a data-driven weighting model derived empirically in a user study involving 151 users of the Data Access Portal (DAP) repository deployed by CSIRO [5]. The recommendation approach is evaluated in another user study that considers more than 1,000 explicit relevance judgments provided by 113 DAP users. The results of the study demonstrate that our approach is capable of accurately predicting the relevance scores of datasets that align with the user judgments. Hence, this is an important step towards practical deployment of recommendation technologies in research data repositories.

In summary, the main contributions of this work are two-fold. First, we propose a new recommendation approach that hybridizes established recommendation techniques to address a *new problem of open research dataset discovery*. Second, we apply the proposed approach in the context of a live dataset repository and *evaluate its accuracy in a user study* that involves dataset creators and consumers alike.

UMAP '18, July 8–11, 2018, Singapore, Singapore

DOI: https://doi.org/10.1145/3209219.3209250

[1] http://www.re3data.org/
[2] https://search.datacite.org/

2 RELATED WORK

Recommender systems have been recognized as an important feature in the design of scientific digital libraries [2, 19]. However, existing work has primarily focused on recommendable items like publications [7], scholars [25] and citations [10]. Previous studies have also indicated that these method may be insufficient for research dataset discovery purposes, as users may have different requirements and employ a range of strategies searching for datasets compared to publication-related discovery [8, 12–14, 20, 22]. For example, the representation of datasets is more complex than that of publications [20, 22], serendipity and diversity are important considerations in dataset discovery [1, 8], and users also appreciate more the novelty of recommendations [12].

To the best of our knowledge, little work has focused on research dataset recommendations. A notable exception to this is the work of Singhal *et al.* on context-based search for research datasets [21]. There, the authors employed content-based similarity based on three features – topic, abstract, and author – in order to identify datasets of interest. Catania *et al.* demonstrated a prototype recommendation approach for spatial data based on the topological similarities of spatial objects [3]. Although their work is applicable to the problem of discovering spatial datasets by considering fine-grained geometrical information, such an approach may be less practical if applied to datasets in open data repositories, as these usually contain only coarse spatial information like bounding boxes and points.

3 DATASET RECOMMENDATION

Our approach combines content-based similarity [16] with item-to-item co-occurrence [15]. The former quantifies the similarity of datasets by comparing their metadata, e.g., keywords and fields of research, while the latter considers their statistical co-occurrence, e.g., downloads by the same users.

3.1 Data Sources

We distinguish between two types of information associated with open research datasets. The first is the metadata of the datasets, as specified by data creators or providers (see Figure 1). The metadata is usually represented using standard schemas, such as Dublin Core[3], DataCite Metadata Schema[4] and Registry Interchange Format - Collections and Services[5]. The second contains the observable user interactions with datasets and the repository. For example, past searches and dataset downloads may be extracted from the repository logs and these reveal commonalities or behavioral patterns of the users. Typically, this information is more abundant than metadata and can be used for dataset recommendations.

Based on the metadata of the DAP repository and the available repository logs, we identified 10 features to be exploited by the proposed recommendation approach: *title, description, keywords, activity, research fields, creators, contributors, spatial, search* and *download*. A brief explanation of these features is provided in Table 1. Note that the first eight features belong to metadata, while the last two are derived from user interaction logs.

[3]http://dublincore.org/documents/dces/
[4]https://schema.datacite.org/
[5]http://www.ands.org.au/online-services/rif-cs-schema

Figure 1: An example of a research dataset and its metadata.

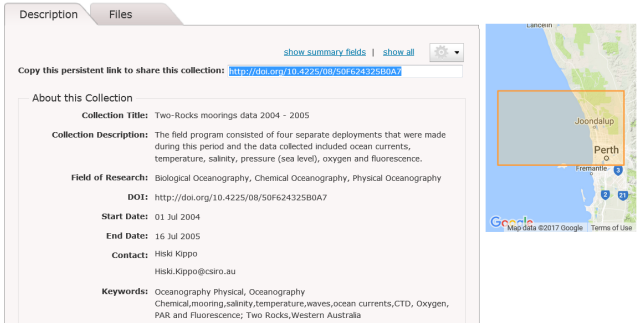


Table 1: Features, descriptions and weights.

Feature	Description	Score	ω_i
title	Name of the dataset	4.106	0.123
description	Textual description of the data	3.887	0.116
keywords	User-specified tags	3.815	0.114
activity	Data provenance: related project, methods, experiments, instruments	3.311	0.099
research fields	Research classification areas	2.669	0.080
creators	Users who created the dataset	2.868	0.086
contributors	Users who contributed to the dataset	2.589	0.077
spatial	Spatial location of the data collected, e.g., point or bounding box	3.523	0.105
download	Downloading-related server logs	N/A	0.100
search	Search-related server logs	N/A	0.100

3.2 Feature-Based Similarity

The proposed dataset recommendation task considers the following use case. We assume that a DAP repository user is examining a target dataset d, e.g., sample dataset shown in Figure 1, and would like to be recommended a list of n related datasets $(d_1, d_2, \ldots, d_n)$. This is similar to the familiar "*users who buy this product are also interested in*" recommendation paradigm, deployed by eCommerce sites. Note that this list of recommended datasets should be ranked according to their relevance to d, quantified using a similarity function $overall\_sim(d, d_k)$, such that d_1 is the most similar and d_n is the least similar dataset.

We compute the similarity of the datasets using a linearly weighted hybridization of two methods. **Content-based** (CB) similarity component determines similar datasets based on the eight metadata features: *title, description, keywords, activity, research fields, creators, contributors* and *spatial*. For the textual features *title, description, keywords* and *activity*, we conduct standard text pre-processing steps, such as stop-word removal, tokenization and stemming, and then use the TF-IDF term weighting and Cosine Similarity to compute the similarity score for each feature [17].

For the categorial *research fields, creators* and *contributors* features, we apply the Jaccard similarity coefficient to compute feature-specific similarity scores. The *spatial* information of research datasets is expressed as a point or bounding box. We transform the spatial information into a standardized geographic coordinate system, compute the centroids of the bounding boxes and then apply the Euclidean distance to compute the spatial distance between datasets. We also normalize the obtained distance matrix and convert it into the similarity matrix.

We also use **item-to-item** (I2I) dataset co-occurrence similarity. This is based on the frequency of joint downloads by DAP users extracted from the *download logs* feature. The underlying idea is that the more two datasets are downloaded jointly by users, the more likely they are to be similar. From the DAP download logs, we extract the associations between datasets and users based on the observed download activity. Then, we represent each dataset as a vector expressing the number of downloads by every user and compute the download-based dataset-to-dataset similarity by applying Cosine Similarity to the two vectors.

To compute similarity using the *search logs* feature, we uncover the relations between datasets and search terms from the DAP logs. To this end, we are able to track which datasets were examined by users from the results[6] returned by DAP in response to search queries. The underlying assumption here is that two datasets are related if they are examined by users after launching similar queries. Hence, we extract the relations between the queries and examined datasets and then compute the search-based dataset-to-dataset similarity by applying Cosine Similarity to the vectors representing the queries.

3.3 Relevance Ranking

In order to compute the overall dataset-to-dataset similarity score $overall\_sim(d, d_k)$, we combine the 10 individual feature-based similarity scores in a linear manner. More formally, the overall similarity of datasets d and d_k is computed by

$$overall\_sim(d, d_k) = \sum_i (\omega_i \cdot sim_i(d, d_k)), \quad (1)$$

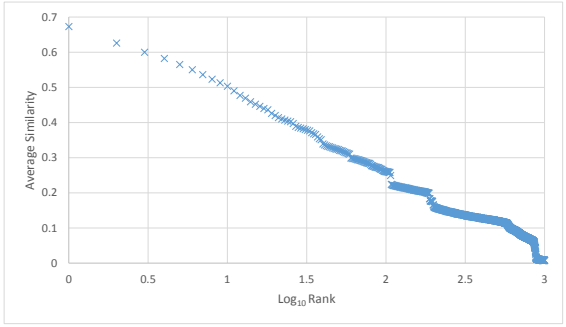
where ω_i refers to the weight associated with a feature i and $sim_i(d, d_k)$ is the similarity of d and d_k with respect to i.

The features specified in Section 3.2 may have different levels of importance for discovering relevant datasets. We use a heuristic weighting model that assigns 0.8 of the weight to the CB similarity score computed using the metadata features and the remaining 0.2 to the I2I co-occurrence similarity. This is in line with the previously used heuristic weights assigned in [11].

Individual weights associated with the eight metadata features were determined empirically through a user study involving the users of DAP [5]. The users were shown all the features and asked to rate on a 5-Likert scale the perceived importance of the features. 151 users provided their ratings and the average scores of the metadata features are shown in Table 1. The survey reveals that *title, description* and *keywords* are deemed to be the more important features scoring closely to 4, whereas *creators, contributors* and *resarch fields* are less important. The obtained importance scores informed the weights of the features, ω_i, which are listed in the right column in Table 1. The weights of the metadata features were normalized, such that $\sum_i \omega_i = 0.8$, as per our heuristic assignment. The weights of the interaction-based *search logs* and *download logs* features were uniformly split to $\omega_i = 0.1$ each.

[6]We disregard the search mechanism of DAP and treat is a "deterministic black box".

Figure 2: Average similarity of 1000 most similar datasets.



Finally, we compile the recommendation lists and include there the datasets with the highest similarity $overall\_sim(d, d_i)$. More precisely, we compute offline the dataset-to-dataset similarity matrix and consider n most similar entries of a target dataset d to be the recommended datasets of relevance.

4 EVALUATION

This section presents the experimental setting and the results obtained in our user study.

4.1 Experimental Setting

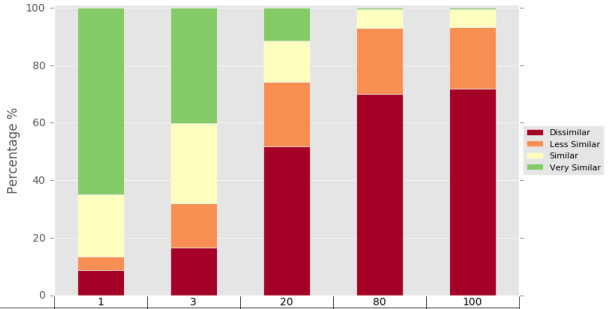
DAP is a data repository deployed by CSIRO, which provides access to datasets published by researchers across various disciplines. In this study, we used 1877 datasets published between 2011 and 2017. We retrieved their metadata and extracted the search and download logs of DAP . After cleansing, the search logs contain 58K unique queries submitted by 13.5K users between 2012 and 2014. The download logs contain more than 10K dataset downloads performed by 6.5K users between 2012 and 2016.

For each of the 1877 datasets in the evaluation, we applied the similarity function given in Equation 1 to compute the similarity scores of all the other datasets in DAP and selected 1000 most relevant datasets. The average $overall\_sim(d, d_k)$ at rank $k \in [1, 1000]$ computed across all the datasets is shown in Figure 2. As can be seen, the similarity scores exhibit a power law distribution. Only the top 10 datasets have $overall\_sim(d, d_k) > 0.5$, whereas datasets ranked 40 and on all have $overall\_sim(d, d_k) < 0.35$.

DAP datasets are published through predefined data descriptors, which leads to consistent nomenclature of the datasets. Hence, it is difficult to accurately establish the ground truth relevance of datasets and evaluate either predictive (MAE, RMSE, etc), classification (precision, recall, etc), or ranking (NDCG, MRR, etc) accuracy metrics [9]. Due to this, we assessed the ability of our approach to predict the relevance of the recommended datasets in a user study involving real DAP users.

We invited a selection of active DAP users to trial a new dataset recommendation feature. For each of the 1877 datasets, we computed offline the list of 1000 most relevant datasets, selected five

Figure 3: Distribution of relevance judgments.



	1	3	20	80	100
Dissimilar	8.8	16.7	51.9	69.9	71.8
Less Similar	4.6	15.3	22.2	23.1	21.3
Similar	21.8	27.8	14.4	6.5	6.5
Very Similar	64.8	40.3	11.6	0.5	0.5

datasets at fixed[7] ranks $k = 1, 3, 20, 80, 100$ and showed these to users as "similar datasets". The users were asked to rate the relevance of the five recommended datasets on a 4-Likert scale ranging from 'very similar' to 'dissimilar'. The datasets were displayed in a random order to avoid selection bias, were visualized in the same way and included DAP links allowing users to inspect them. Users could also include free-text feedback justifying their ratings.

4.2 Results

113 DAP users participate in the second study and evaluated 216 target datasets, i.e., five "similar datasets" for each, such that, in total, we obtained 1080 explicit relevance judgments. Figure 2 illustrates the distribution of the relevance judgements assigned by the users. The five bars represent the distributions of user judgments at ranks 1, 3, 20, 80, and 100, respectively.

Overall, more than 86% of the datasets at rank 1 were judged to be either 'relevant' or 'highly relevant'. Notably, less than 9% of the datasets at this rank were rated as 'dissimilar'. This shows that the proposed approach in most cases can accurately predict the relevance of the top-ranked dataset. Inspecting the free-text feedback provided by the users, we discovered that they rated these datasets as relevant due to a range of reasons: target dataset and the recommended dataset were thematically related, were generated by the same project or using a similar method, formed a series of measurements or were just derived one from another.

For example, one user wrote: "*[The target dataset] was an input to [the recommended dataset]. I consider this to be very similar or strongly related*". Another user commented: "*[The target dataset] and [the recommended dataset] are soil property predictions for the project. They have different attributes, but are similar because users usually look for a suite of soil attributes covering the same area*".

Similarly, out of all the datasets at rank 3, more than 68% were judged as 'relevant' or 'highly relevant' and less than 17% were rated as 'dissimilar'. Although positive sentiment toward datasets at rank 3 still dominates, the number of negative judgments is almost twice larger than for datasets at rank 1. This observation shows that the proposed approach is sensitive enough to differentiate even between datasets at ranks as high as 1 and 3, and this is evident in the users' relevance judgments.

The situation is different, however, at ranks 20, 80 and 100, where, respectively, close to 74%, 93% and 93% of the obtained judgements rated the recommended datasets as either 'dissimilar' or 'less similar'. In fact, results obtained for ranks 80 and 100 were very close and at both ranks only 7% of the datasets were judged to be 'relevant' or 'highly relevant'. This aligns with the similarity distribution observed in Figure 2, as both the ranks are located at the tail of the distribution. Although the results at rank 20 were slightly better than at 80 and 100, only 26% of the datasets at this rank were rated as 'relevant' or 'highly relevant'. This shows that the users perceive the moderately-ranked datasets to be mostly irrelevant for the recommendation purposes.

Interestingly, only one dataset at rank 80 was rated as 'very similar' and this happened to be collected by the same project as the target dataset. Also among the datasets at rank 100, only one dataset was rated as 'very similar'. The user who positively evaluated this dataset commented: "*[The recommended dataset] contains the sea temperatures measured on a vessel as it steams underway. [The target dataset] is also a sea temperature dataset taken by a vessel. Because both datasets have the same measured parameter with the same temporal and spatial attributes, I classified them as similar, albeit [the recommended dataset] has many other parameters. Someone searching for sea temperature data could use that data from [the recommended dataset]*".

5 CONCLUSIONS AND FUTURE WORK

The proliferation of open research datasets may aggravate the discovery of datasets-of-interest by users. In order to address this growing issue, we have turned in this paper to the problem of recommending datasets. We developed a recommendation approach that identifies relevant datasets by leveraging a hybrid similarity metric, which incorporates content-based metadata features and observable usage patterns. Notably, the linear weighting model deployed by our approach was derived in a user study involving the users of a real DAP open dataset repository.

We evaluated the proposed approach in another user study that is reported in this paper. The results of the latter verified the ability of the proposed approach to accurately predict the relevance of datasets, which we consider to be an important contribution to the challenge of recommending research datasets. The obtained relevance judgements of DAP users indicate that the datasets recommended at ranks 1 and 3 were mainly deemed as relevant, whereas those at ranks 20, 80 and 100 were clearly irrelevant.

The study addresses a novel recommendation problem, and its results may serve as baseline for future research in this direction. In particular, we will measure the performance of three methods, i.e., content-based, co-occurrence and hybrid proposed in the paper. We intend to compare the performance of our approach with established content-based recommenders from other domains, deployed for dataset recommendations. We will then conduct a large-scale live evaluation of the recommender and assess its uptake by the DAP users.

[7] These ranks were picked based on the distribution of the average similarity scores (see Figure 2) and represent different relevance levels. Specifically, dataset at ranks 1 and 3 have $overall\_sim(d, d_k) > 0.6$ and are expected to be relevant, whereas those at ranks 80 and 100 have $overall\_sim(d, d_k) < 0.3$ and should be irrelevant.

REFERENCES

[1] Azin Ashkan, Branislav Kveton, Shlomo Berkovsky, and Zheng Wen. 2015. Optimal Greedy Diversity for Recommendation. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*. 1742–1748. http://ijcai.org/Abstract/15/248

[2] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries* 17, 4 (2016), 305–338. https://doi.org/10.1007/s00799-015-0156-0

[3] Barbara Catania, Maria Teresa Pinto, Paola Podestà, and Davide Pomerano. 2011. *A Recommendation Technique for Spatial Data*. Springer Berlin Heidelberg, Berlin, Heidelberg, 200–213. https://doi.org/10.1007/978-3-642-23737-9_15

[4] Miriam L. E. Steiner Davis, Carol Tenopir, Suzie Allard, and Michael T. Frame. 2014. Facilitating Access to Biodiversity Information: A Survey of Users' Needs and Practices. *Environmental Management* 53, 3 (01 Mar 2014), 690–701. https://doi.org/10.1007/s00267-014-0229-7

[5] Anusuriya Devaraju and Shlomo Berkovsky. 2017. Do Users Matter? The Contribution of User-Driven Feature Weights to Open Dataset Recommendations. In *Proceedings of the Poster Track of the 11th ACM Conference on Recommender Systems (RecSys 2017), Como, Italy, August 28, 2017*. http://ceur-ws.org/Vol-1905/recsys2017_poster16.pdf

[6] Ixchel M. Faniel, Adam Kriesberg, and Elizabeth Yakel. 2016. Social scientists' satisfaction with data reuse. *Journal of the Association for Information Science and Technology* 67, 6 (2016), 1404–1416. https://doi.org/10.1002/asi.23480

[7] C. Lee Giles, Kurt D. Bollacker, and Steve Lawrence. 1998. CiteSeer: An Automatic Citation Indexing System. In *Proceedings of the Third ACM Conference on Digital Libraries (DL '98)*. ACM, New York, NY, USA, 89–98. https://doi.org/10.1145/276675.276685

[8] Kathleen Gregory, Paul T. Groth, Helena Cousijn, Andrea Scharnhorst, and Sally Wyatt. 2017. Searching Data: A Review of Observational Data Retrieval Practices. *CoRR* abs/1707.06937 (2017). http://arxiv.org/abs/1707.06937

[9] Asela Gunawardana and Guy Shani. 2015. Evaluating Recommender Systems. In *Recommender Systems Handbook*. 265–308. https://doi.org/10.1007/978-1-4899-7637-6_8

[10] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware Citation Recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW '10)*. ACM, New York, NY, USA, 421–430. https://doi.org/10.1145/1772690.1772734

[11] Pijitra Jomsri, Siripun Sanguansintukul, and Worasit Choochaiwattana. 2011. CiteRank: combination similarity and static ranking with research paper searching. *International Journal of Internet Technology and Secured Transactions* 3, 2 (2011), 161–177. http://www.inderscienceonline.com/doi/abs/10.1504/IJITST.2011.039776

[12] Dagmar Kern and Brigitte Mathiak. 2015. *Are there any differences in data set retrieval compared to well-known literature retrieval?* Springer International Publishing, Cham, 197–208. https://doi.org/10.1007/978-3-319-24592-8_15

[13] Sven R. Kunze and Sören Auer. 2013. Dataset Retrieval. In *2013 IEEE Seventh International Conference on Semantic Computing, Irvine, CA, USA, September 16-18, 2013*. 1–8. https://doi.org/10.1109/ICSC.2013.12

[14] Branislav Kveton and Shlomo Berkovsky. 2016. Minimal Interaction Content Discovery in Recommender Systems. *TiiS* 6, 2 (2016), 15:1–15:25. https://doi.org/10.1145/2845090

[15] Loet Leydesdorff and Liwen Vaughan. 2006. Co-occurrence Matrices and Their Applications in Information Science: Extending ACA to the Web Environment. *J. Am. Soc. Inf. Sci. Technol.* 57, 12 (Oct. 2006), 1616–1628. https://doi.org/10.1002/asi.v57:12

[16] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. 2011. *Content-based Recommender Systems: State of the Art and Trends*. Springer US, Boston, MA, 73–105. https://doi.org/10.1007/978-0-387-85820-3_3

[17] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA. https://nlp.stanford.edu/IR-book/html/htmledition/contents-1.html

[18] Jennifer C Molloy. 2011. The Open Knowledge Foundation: open data means better science. *PLoS biology* 9, 12 (Dec. 2011), e1001195. https://doi.org/10.1371/journal.pbio.1001195

[19] D. De Nart and C. Tasso. 2014. A Personalized Concept-driven Recommender System for Scientific Libraries. *Procedia Computer Science* 38 (2014), 84 – 91. https://doi.org/10.1016/j.procs.2014.10.015 10th Italian Research Conference on Digital Libraries, IRCDL 2014.

[20] S. L. Pallickara, S. Pallickara, M. Zupanski, and S. Sullivan. 2010. Efficient Metadata Generation to Enable Interactive Data Discovery over Large-Scale Scientific Data Collections. In *2010 IEEE Second International Conference on Cloud Computing Technology and Science*. 573–580. https://doi.org/10.1109/CloudCom.2010.99

[21] A. Singhal, R. Kasturi, and J. Srivastava. 2014. DataGopher: Context-based search for research datasets. In *Proceedings of the 2014 IEEE 15th International Conference on Information Reuse and Integration (IEEE IRI 2014)*. 749–756. https://doi.org/10.1109/IRI.2014.7051964

[22] Maximilian Stempfhuber and Benjamin Zapilko. 2009. Integrated Retrieval of Research Data and Publications in Digital Libraries. In *Rethinking Electronic Publishing: Innovation in Communication Paradigms and Technologies - Proceedings of the 13th International Conference on Electronic Publishing*, Susanna Mornati and Turid Hedlund (Eds.). 613–620. http://elpub.scix.net/cgi-bin/works/Show?144_elpub2009

[23] Christine Stohn. 2015. *How Do Users Search and Discover? Findings from Ex Libris User Research*. Technical Report. Ex Libris. http://www.exlibrisgroup.com/files/Products/Primo/HowDoUsersSearchandDiscover.pdf

[24] Ryen W. White and Resa A. Roth. 2009. *Exploratory Search: Beyond the Query-Response Paradigm*. Morgan & Claypool Publishers. http://dx.doi.org/10.2200/S00174ED1V01Y200901ICR003

[25] A. Yang, J. Li, Y. Tang, J. Wang, and Y. Zhao. 2012. The similar scholar recommendation in Schol@t. In *Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. 666–670. https://doi.org/10.1109/CSCWD.2012.6221889