Sociometric methods for relevancy analysis of Long Tail Science Data

Arcot Rajasekar, Sharlini Sankaran,
Howard Lander, Tom Carsey, Jonathan Crabtree
The University of North Carolina Chapel Hill
Chapel Hill, North Carolina, USA.
{rajasekar, sharlini, howard_lander, carsey,
jonathan_crabtree, }@unc.edu

Merce Crosas, Gary King
Harvard University
Cambridge, Massachusetts, USA.
{mcrosas, king}@harvard.edu

Hye-Chung Kum
Texas A&M University
College Station, Texas, USA.
kum@srph.tamhsc.edu

Justin Zhan
North Carolina A&T State University
Greensboro, North Carolina, USA.
zzhan@ncat.edu

Abstract— **As the push towards electronic storage, publication, curation, and discoverability of research data collected in multiple research domains has grown, so too have the massive numbers of small to medium datasets that are highly distributed and not easily discoverable – a region of data that is sometimes referred to as the long tail of science. The rapidly increasing, sheer volume of these long tail data present one aspect of the Big Data problem: how does one more easily access, discover, use, and reuse long tail data to lead to new multidisciplinary collaborative research and scientific advancement? In this paper, we describe DataBridge, a new e-science collaboration environment that will realize the potential of long tail data by implementing algorithms and tools to more easily enable data discoverability and reuse. DataBridge will define different types of semantic bridges that link diverse datasets by applying a set of sociometric network analysis (SNA) and relevance algorithms. We will measure relevancy by examining different ways datasets can be related to each other: data to data, user to data, and method to data connections. Through analysis of metadata and ontology, by pattern analysis and feature extraction, through usage tools and models, and via human connections, DataBridge will create an environment for long tail data that is greater than the sum of its parts. In the project's initial phase, we will test and validate the new tools with real-world data contained in the Dataverse Network, the largest social science data repository. In this short paper, we discuss the background and vision for the DataBridge project, and present an introduction to the proposed SNA algorithms and analytical tools that are relevant for discoverability of long tail science data.**

Keywords- Long tail data, sociometric network analysis, data discoverability

I. INTRODUCTION

With the internet celebrating over a quarter century of existence, electronic storage and publication of data has become almost second nature to most scientists. In the past few years, the push towards open access and discoverability of research data has grown, as evinced by a White House Office of Science and Technology Policy memo addressing open access to data and directing federal agencies to provide funding towards initiatives that increase accessibility of publicly-funded research [1]. Contemporaneously, there has been a rapid growth in the large numbers of small datasets that are highly distributed, not well organized or curated, and thus are not easily discoverable or reusable. These datasets typically exist in total isolation from each other, are individually managed, and suffer from sparse and inconsistent provenance and metadata. Palmer et al. [2, 3] refer to these as the "long tail of science" data – the massive number of relatively small datasets which currently make up the largest proportion of scientific research data.

Though these datasets are small and individually easy to manage, they contain rich information that can be used and reused to maximize new scientific discoveries – *if* the data are easily discoverable, accessible, and analyzable. Discovering long tail data is hard because the data is often distributed in personal workspaces with little attempt made at data publication. Finding *relevant* data is made even more problematic by the difficulty in defining relevancy metrics for scientific datasets. Accessing relevant data is not easy when the data are distributed, not well documented, and in heterogeneous and possibly unique formats. These same characteristics also inhibit the analysis of data and are a major obstacle to generating new knowledge from this type of research data.

While we often use the term 'Big Data' to refer to very large datasets that pose problems in management and analysis due to their sheer size, these only represent one aspect of the Big Data problem. The rapidly-growing amounts of long tail data pose a different, yet fundamental, Big Data problem: how do we enable easier discoverability, use, and reuse of the massive number of smaller datasets that exist in almost total isolation from one another? How do we determine what datasets are relevant to others, and thus make discoverability of these relevant data easier? In this paper we discuss the **DataBridge** [4, 5], a collaborative tool to address some of the challenges associated with long tail data.

DataBridge is an NSF-funded collaboration between University of North Carolina at Chapel Hill, Harvard University, and North Carolina A&T State University to develop an e-Science environment that measures relationships between different datasets. In the DataBridge, similarity and relevancy of long tail data will be assessed on four general aspects: the contents of the data itself, contextual information about the data, producers and consumers of the data, and methods used to create and analyze the data. Using these relationships, we can discover and maintain profiles and clusters for datasets to help researchers seek, search, browse and identify data relevant to their work.

In the next section of this paper, we discuss the overarching vision for the design of the DataBridge. Section III of the paper further discusses the concept of sociometric network analysis (SNA) and details various methods used to measure the relevance relationships among datasets. Section IV presents related works, with a conclusion in Section V.

II. DATABRIDGE VISION

DataBridge is an indexing mechanism for scientific datasets, similar to web search engines that help find web pages of interest. Unlike web search engines that use the textual content of a web page and hyperlinks to identify its relevance to a query, the search space for scientific datasets is quite different and needs external resources such as tags, metadata, contexts, and naming conventions to identify relevancy. As discussed in the introduction, typical long tail datasets in isolation provide very sparse information content for search and discovery.

A resource discovery system for scientific datasets should provide a rich set of tools for mining information and context. To this end, the DataBridge system will analyze linkages between datasets. It will gather data, metadata, usage and other information, and apply SNA algorithms to map datasets connected by multi-dimensional relationships. In this multi-dimensional network, sub-graphs, clusters, and cliques will be used to inform the discovery of other relevant datasets.

Even though a large number of datasets still remain only stored in personal workspaces, without formal organization and metadata, successful efforts have been made to provide centralized data repositories to properly share, manage and archive scientific datasets. These efforts include the Dataverse Network (DVN) and the Integrated Rule-Oriented Data System (iRODS). In the initial phase, DataBridge will draw upon these existing systems because they offer a rich set of real-world structured data and metadata that will help validate the algorithms and analysis. The DVN is an e-Science collaboration environment used to publish, share, reference, extract and analyze research data [6, 7, 8, 9, 32]. Together, University of North Carolina and Harvard University host DVN instances containing over 50,000 research studies with more than 700,000 data and supplementary files in social science. We chose the DVN as a starting point because of its richness of data, but also because the DVN facilitates long term access and good data archival practices while the researcher retains control of, and recognition for, the data he or she deposits. iRODS is a data grid middleware [10, 11, 12, 13, 14, 15] which provides many facilities for collection building, managing, querying, accessing, and preserving data in a distributed data grid framework. The iRODS system applies policy-based control when performing these functions.

DataBridge will eventually gather information from multiple data resources maintained by individuals, projects, regional or disciplinary repositories, and national collaboratives. This information will be integrated into a semantically-rich interface that will allow discoverability of relevant datasets based on relationships between data, users, methods and metadata. Internally, it will apply several relevance algorithms, described in the next section, to build a knowledge base and generate interconnected networks. The DataBridge will provide a venue for scientists to publish, discover, and access data of importance and to find others engaged in similar and pertinent research. The resulting networks and additional information gathered by DataBridge will be transferred back to permanent data repositories such as DVN, to facilitate discoverability within the repository and automatically enhanced the curation of datasets.

Given a set of criteria such as a sample dataset, DataBridge will query its knowledge base to find relevant datasets that are close to the initial dataset based on the given criteria. To do this, the architecture of DataBridge will consist of three functional units: a data gatherer which interacts with data repositories to gather information about scientific datasets, a relevance engine which integrates information about datasets into a relevance network based on sociometric analysis, and a web-based user interface which searches for relevant datasets and gathers information through crowd sourcing and collaborative tagging.

In the following section, we describe the proposed SNA and relevancy measures to be applied in the DataBridge.

III. SOCIOMETRIC NETWORK ANALYSIS FOR LONG TAIL SCIENCE DATA

The DataBridge effort will create a semantically-rich, cross-disciplinary, sociometric network using modern information network analysis tools that build on seminal work by Jacob L. Moreno [16]. Moreno described sociometry as "the inquiry into the evolution and organization of groups and the position of individuals within them." Moreno pioneered the depiction of social relationships between people through sociograms – graphs that symbolize individuals as nodes connected by links, or edges. DataBridge will be the first attempt to apply sociometry and its derived techniques to the study of scientific datasets.

A. Sociometric Network Analysis Algorithms

Successful sociometric algorithms hinge upon adequately detecting community structure, or clustering, in real systems. Though many disciplines, including sociology, biology, and computer science, often represent individuals as graphs, how to detect 'community' has not yet been satisfactorily solved, despite the huge effort [17, 18, 19] of a large interdisciplinary community of scientists over the past few years. Building upon the efforts of this previous work, we will explore a broad range of community detection methods for DataBridge.

The main elements of the problem of community detection in graphs are not defined – indeed, there is not a current, universally accepted definition of 'community' [17, 18, 19]. For the purposes of graph clustering, we define communities as groups of nodes similar to each other. Similarity between each pair of nodes can be computed with respect to a previously assigned reference property, whether or not the nodes are connected. Each node is assigned to the cluster of nodes most similar to it. If the graph nodes are embedded in an n-dimensional Euclidean space, the distance between a pair of nodes can be used as a measure of similarity. For the design of the Databridge, we will investigate several algorithms to quantify clustering including: relative proximity, cosine similarity, dissimilarity, random walk, and resistance distance measures.

The relative proximity between any two data points A=(a_1,a_2, ⋯ ,a_n) and B=(b_1,b_2, ⋯ ,b_n) can be measured by any norm L_m, such as the Euclidean distance (L_2- norm [SM1]), the Manhattan distance (L_1-norm [SM2]), or the L_∞-norm [SM3]. Another popular spatial measure is the cosine similarity [SM4]. If the graph cannot be embedded in space, similarity must be inferred from the adjacency relationships between nodes. Another possibility based on the concept of structural equivalence [SM5] [32] defines the distance between nodes [20] as a dissimilarity measure: two nodes are structurally equivalent if they have the same neighbors, even if they are not adjacent themselves. Using the dissimilarity measure, d_ij=0 if i and j are structurally equivalent. Nodes with large degree and different neighbors are considered very "far" from each other and will have a greater dissimilarity measure. Alternatively, one could measure the overlap between the neighborhoods Γ(i) and Γ(j) of vertices i and j, given by the ratio between the intersection and the union of the neighborhoods [SM6]. Another measure related to structural equivalence is the Pearson correlation [SM7] between columns or rows of the adjacency matrix. Table 1 is a summary of these measures.

An alternative measure is the number of edge- (or node-) independent paths between two nodes. Independent paths do not share any edge (node), and their number is related to the maximum flow that can be conveyed between the two nodes under the constraint that each edge can carry only one unit of flow (max-flow/min-cut theorem) [21]. The maximum flow can be calculated in a time O(m), for a graph with m edges, using techniques like the augmenting path algorithm. Similarly, one could consider all paths running between two nodes. In this case, there is the problem that the total number of paths is infinite, but this can be avoided if one performs a weighted sum of the number of paths. For instance, paths of length L can be weighted by the factor α^l, with α<1. Another possibility, suggested by Estrada and Hatano [22, 23], is to weigh paths of length L with the inverse factorial 1/l!. In both cases, the contribution of long paths is strongly suppressed and the sum converges.

Random Walk properties provide another class of measures of node similarity that can be used in sociometric network analysis of long tail data. For example, one random walk property measures the commute time between a pair of nodes, or the average number of steps needed for a random walker, starting at either node, to reach the other node for the first time and to come back to the starting node.

The commute time and various variants have been used as a similarity measure by Saerens and coworkers [24, 25, 26, 27]: the larger the time, the farther (less similar) the nodes. The commute time [28] is closely related to another measure, the resistance distance [29]. The resistance distance expresses the effective electrical resistance between two nodes if the graph were turned into a resistor network. White and Smyth [18, 30] used instead the average first passage time, i.e. the average number of steps needed to reach for the first time the target node from the source.

Table 1. Similarity Measures (SM)

SM1	The Euclidean distance (L_2-norm)	$d^{E}_{AB} = \sum_{k=1}^{n} \sqrt{(a_k - b_k)^2}$
SM2	The Manhattan distance (L_1-norm)	$d^{M}_{AB} = \sum_{k=1}^{n} \lvert a_k - b_k \rvert$
SM3	The L_∞-norm	$L_1 d^{\infty}_{AB} = \max_{k\epsilon[1,n]} \lvert a_k - b_k \rvert$
SM4	Cosine similarity Range of ρ_{AB} is [0,π)	$\rho_{AB} = arccos \frac{a \cdot b}{\sqrt{\sum_{k=1}^{n} a_k^2}\sqrt{\sum_{k=1}^{n} b_k^2}}$, where $\boldsymbol{a} \cdot \boldsymbol{b}$ is the dot product of the vectors **a** and **b.**
SM5	Dissimilarity measure in structural equivalence	$d_{ij} = \sqrt{\sum_{k\neq i,j}(A_{ik} - A_{jk})^2}$, where $\boldsymbol{A}$ is the adjacency matrix**.**
SM6	Overlap in neighborhood	$\omega_{ij} = \frac{\lvert \Gamma(i) \cap \Gamma(j) \rvert}{\lvert \Gamma(i) \cup \Gamma(j) \rvert}$, neighborhoods Γ(i) and Γ(j) of vertices i and j
SM7	Pearson correlation between columns or rows	$C_{ij} = \frac{\sum_k (A_{ik} - \mu_i)(A_{jk} - \mu_j)}{n\sigma_i \sigma_j}$, where the averages $\mu_i = (\sum_j A_{ij})/n$ and the variances $\sigma_i = \sqrt{\sum_j (A_{ij} - \mu_i)^2/n}$.

B. Relevance Algorithms:

In order to connect disparate datasets in a network and discover multidimensional similarities, we will research and implement several types of relevance metrics. These metrics can be broadly grouped into: data to data connections, user to data connections, and method to data connections, as described below.

1) Data to Data Connections: The ability to understand and connect data across data types and research disciplines hinges on the quality of metadata available to describe these data. The current process of creating metadata is very labor- and time- intensive. Many long tail data suffer from sparse metadata because archives and repositories are forced to make a tradeoff to process many datasets with minimum description or to process fewer datasets and take time to create more detailed metadata. The DataBridge will include a service that generates metadata based upon the results of relevance engine processing. This new interface will automatically suggest topics and keywords to researchers uploading and ingesting data into repositories by pre-tagging the projected topic space with appropriate terms. Accurate suggestions will encourage accurate crowd sourced tagging of the ingested data. This researcher guided/machine-learning metadata creation environment will have a profound impact on the future of data discovery and reuse [31].

As we build the data to data relevancy engine, we will compare different probabilistic topic models to determine which work best to cluster a corpus of short abstracts, explore effective metrics for sampling the abstracts to be projected together using these models, and methods to effectively scale these models (i.e. layer the topic space). We will further compare the effectiveness of these methods to random projections, which can scale to handle a large corpus.

Semantic similarity of datasets can be measured by applying technologies in Natural Language Processing (NLP) to infer the topic space as latent classes. Methods such as Probabilistic Latent Semantic Analysis (PLSA) can effectively account for both synonyms (words that refer to the same topic) and polysemy (words with multiple meanings) in a corpus of abstracts by modeling each document as a mixture of topics, each being a unigram model.

2) User to Data Connections: Using the DVN as a starting point for modeling user to data connections, we are able to search metadata fields including authors, producers, distributors, provenance, and geographic and time coverage of a dataset. DataBridge will extend beyond the current manual, passive search options and will allow for better understanding of and even prediction of possible future collaborations. We will crawl published papers that use DVN data to identify datasets from past collaborations and explore collaboration patterns along features in the given datasets. Using these patterns and some of the similarity measures discussed in Section III B, we will build models of collaboration to predict data connections.

3) Methods to Data Connections: The use of particular models and methods to analyze datasets provides rich data for mining sociometric information. Usage methods and applications can be viewed as properties of the datasets and can be used to determine relevance between datasets. Since scant research has been conducted on measures of similarity between research methods and long tail data, an ontology of methods, tools, and applications needs to be defined. From this information, we plan to implement relevance algorithms that will use the method ontology to help define a relevance network.

4) Interactive Connections: Some of us (King and Crosas together with Grimmer, Stewart and members of the Harvard's IQSS software team) are working on a computer-assisted method to discover clusters (or partitions) in large corpora of unstructured text [46]. This method differs from the ones mentioned above because it allows users to interact with the clustering space until they find a result useful and tuned to their needs.

IV. RELATED WORK

Currently, there are several national consortium-based projects including DataONE (Dataone) [33], Datanet Federation Consortium (DFC) [34], the Consortium of Universities for the Advancement of Hydrologic Science (CUHASI) [35], iPlant Consortium (IPC) [36], and the Ocean Observatories Initiative (OOI) [37] that collect and provide access to disciplinary data collections. Additionally, scientific collaboration tools have been developed that help scientists build collections for their projects. These include systems such as iRODS, the DVN, Fedora Commons (DuraSpace), and LOCKSS (lockss).

What's lacking from these efforts is a network that connects datasets such that the whole becomes greater than the sum of its parts – connections based on similarities beyond normal textual connections within the silos of single disciplines. Missing from these works also is an explicit focus on the relationships among datasets.

Other researchers have begun to explore how to link datasets. Tools such as Scival Experts, CiteSeerX and DataCite, for example, are designed around metadata schemas. SciVal Experts derives relationships between publications and experts by coauthorship and keywords, but does not address datasets per se. CiteSeerX includes some additional techniques such as automatic metadata harvesting from indexed articles and crowd sourcing of opinions about articles, but still does not address datasets. DataCite is designed to make it easier for researchers to find relevant datasets by collecting metadata, providing a search capability, and assigning persistent data identifiers to assist in citing and publication, but it lacks any of the sociometric infrastructure we intend to build in the DataBridge.

In relation to scholarly communication, the role of automatic metadata generation is being researched and tested as a method to help increase discoverability, access, and efficiency [38, 39, 40]. In today's digital information age, many now consider datasets unique units of scholarly communication in their own right [41, 42]. SEAD [43] is also concerned with sustainability of datasets in the long tail of science and proposes to provide a data repository for scientists to manage, share, and link their data with others. However, the linking is done manually by users and not through automatic analysis as proposed here. The Linked Data project [44] links data through the use of a data description language, but the description is not generated by analysis as we propose. Google Scholar allows the user to search a wide variety of sources, including books and some web sites, but has neither a sociometric component nor a focus on datasets. Google does focus on the sociometry of datasets, but these datasets are limited to what are presented on HTTP servers that can be crawled and are normally unstructured text or images. Another related project is the CASRAI program [45], which is focused on developing common metadata standards to allow for linking research information to facilitate data exchange, collaboration, and interaction with funding agencies.

V. CONCLUSION

This paper outlines the background and vision behind DataBridge – an e-Science collaboration system that enables researchers to discover relevant datasets in the long tail of science data by applying SNA algorithms and multiple dimensions of relevancy between datasets. Work is underway in developing this framework and implementing and evaluating the DataBridge system.

ACKNOWLEDGMENT

The efforts described in this paper are funded by the NSF Office of Cyberinfrastructure OCI-1247652, OCI-1247602, OCI-1247663 grant, "BIGDATA: Mid-Scale: ESCE: DCM: Collaborative Research: DataBridge - A Sociometric System for Long tail Science Data Collections", (2012-2015).

REFERENCES

[1] P. Holdren. "Memorandum For The Heads Of Executive Departments And Agencies. SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research," Feb. 2013, http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[2] C. Palmer, M. Cragin, P. Heidorn, and L. Smith, "Data curation for the long tail of science: The case of environmental sciences." Paper presented at the Third International Digital Curation Conference, Washington, DC., Dec. 2007.

[3] C. Palmer and C. Faloutsos, "Electricity based external similarity of categorical attributes," in: Proceedings of PAKDD, 2003, pp. 486-500.

[4] Databridge, http://www.databridge.web.unc.edu

[5] A. Rajasekar, H. Kum, M. Crosas, J. Crabtree, S. Sankaran, H. Lander, T. Carsey, G. King, and J. Zhan. "The DataBridge," Science Journal. ASE in press 2013.

[6] M. Crosas, "The Dataverse Network®: An Open-Source Application for Sharing, Discovering and Preserving Data," D-Lib Magazine, Volume 17 Number 1/2, 2007.

[7] M. Crosas, "A Data Sharing Story," 2012, J eScience Librarianship 1(3): Article 7, 2012.

[8] DVN, The Dataverse Network, http://thedata.org

[9] G. King, "An Introduction to the Dataverse Network as an Infrastructure for Data Sharing, Sociological Methods & Research," vol 36(2), 2007, pp. 173-199.

[10] iRODS, integrated Rule Oriented Data System, http://www.irods.org

[11] R. Moore, and A. Rajasekar, "Rule-Based Distributed Data Management Grid," 2007 IEEE/ACM International Conference on Grid Computing.

[12] R. Moore, A. Rajasekar, A., and A. de Torcy, A.. "Policy-based Digital Library Management," International Conference on Digital Libraries, Delhi, India, Feb. 2009.

[13] A. Rajasekar, M. Wan, R. Moore, and W. Schroeder, "A Prototype Rule-based Distributed Data Management System," HPDC workshop on Next Generation Distributed Data Management, Paris, France, 2006.

[14] A. Rajasekar, R. Moore, R., M. Wan, W. Schroeder, and A. Hasan, "Applying Rules As Policies for Large-Scale Data Sharing," 1st International Conference on Intelligent Systems, Modelling and Simulation, Liverpool, UK., Jan. 2010.

[15] M. Wan, R. Moore, and A. Rajasekar, "Integration of Cloud Storage with Data Grids," The Third International Conference on the Virtual Computing Initiative, Research Triangle Park, NC, Oct. 2009.

[16] J. Moreno, Sociometry, Experimental Method and the Science of Society. An Approach to a New Political Orientation, Beacon House, Beacon, New York, 1951.

[17] R. Gross, and A. Acquisti, "Information revelation and privacy in online social networks," In Pre-proceedings version. ACM Workshop on Privacy in the Electronic Society (WPES), 2005.

[18] F. Lorrain and H. White, "Structural equivalence of individuals in social networks," J. Math. Sociol. Vol. 1 1971, pp. 49-80.

[19] S. Wasserman and K. Faust, Social Network Analysis, Cambridge University Press, Cambridge, UK, 1994.

[20] R. Burt, "Positions in networks," Soc. Forces vol 55, 1976 pp. 93-122.

[21] P. Elias, A. Feinstein, and C.E. Shannon, "Note on maximum flow through a network," IRE Trans. Inf. Theory IT- 2, 1956, pp. 117-119.

[22] E. Estrada and N. Hatano, "Communicability in complex networks," Phys. Rev. E vol. 77 (3), 2008, pp. 036-111.

[23] E. Estrada and N. Hatano, "Communicability graph and community structures in complex networks," Appl. Math. Comput. Vol. 214, 2009, pp. 500-511.

[24] F. Fouss, A. Pirotte, J. Renders, and M. Saerens, "Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation," Knowledge and IEEE Transactions on Data Engineering, , vol.19, no.3, Mar. 2007, pp.355-369.

[25] M. Saerens, F. Fouss, L. Yen, and P. Dupont, "The principal component analysis of a graph and its relationships to spectral clustering," in: Proc. Eur. Conf. on Machine Learning, 2004, citeseer.ist.psu.edu/saerens04principal.html.

[26] L. Yen, F. Fouss, C. Decaestecker, P. Francq, and M. Saerens, "Graph nodes clustering with the sigmoid commute-time kernel: A comparative study," Data Knowl. Eng. Vol. 68 (3), 2009, pp. 338-361.

[27] L. Yen, F. Fouss, C. Decaestecker, P. Francq, and M. Saerens, "Graph nodes clustering based on the commute-time kernel," in: PAKDD, pp. 1037-1045.

[28] A. Chandra, P. Raghavan, W.L. Ruzzo, and R. Smolensky, "The electrical resistance of a graph captures its commute and cover times," in: STOC '89: Proceedings of the twenty-first annual ACM symposium on Theory of computing, ACM, New York, NY, USA, 1989, pp. 574-586.

[29] Klein, D. and Randic, M. (1983). Resistance distance, J. Math. Chem. 12 pp. 81-95.

[30] S. White and P. Smyth, "Algorithms for estimating relative importance in networks," in: KDD '03: Proceedings of the Ninth

ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, NY, USA, 2003, pp. 266-275.

[31] T. Hofmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis,” Machine Learning Journal, 42(1), 2001, pp.177-196.

[32] M. Altman, and G. King, “A proposed standard for the Scholarly Citation of Quantitative Data,” DLib Magazine 13 (3/4). 2007.

[33] Dataone, DataONE, http://www.dataone.org

[34] DFC, Datanet Federation Consortium, http://www.feddata.org

[35] CUHASI, Consortium of Universities for the Advancement of Hydrologic Science, http://www.cuahsi.org/

[36] IPC, The iPlant Collaborative, http://www.iplantcollaborative.org/

[37] Ocean Observatories Initiative (OOI), http://www.oceanobservatories.org/

[38] K. Calhoun, The Changing Nature of the Catalog and its Integration with Other Discovery Tools. Library of Congress. 2006, http://www.loc.gov/catdir/calhoun-report-final.pdf.

[39] J. Greenberg, “Metadata Extraction and Harvesting. Journal of Internet Cataloging,” vol 6(4), 2004, pp. 59–82.

doi:10.1300/J141v06n04_05.

[40] J. Greenberg, H. White, C. Carrier, and R. Scherle, “A Metadata Best Practice for a Scientific Data Repository,” Journal of Library Metadata, vol9(3-4), 2009, pp. 194–212.

http://dx.doi.org/10.7191/jeslib.2012.1020.

[41] H. Davis, and J. Vickery, “Datasets, a Shift in the Currency of Scholarly Communication: Implications for Library Collections and Acquisitions. Serials Review,” vol. 33(1), 2007, pp. 26–32.

doi:10.1016/j.serrev.2006.11.004.

[42] H. Van de Sompel, S. Payette, J. Erickson, C. Lagoze, and S. Warner, “Rethinking scholarly communication: Building the System that Scholars Deserve.” D-Lib Magazine, vol. 10(9), 2004

doi:10.1045/september2004-vandesompel.

[43] SEAD, http://sead-data.net

[44] The Linked Open Data project: http://www.Linkeddata.org

[45] CASRAI: http://casrai.org

[46] Grimmer, Justin, and Gary King. 2011. General Purpose Computer-Assisted Clustering and Conceptualization. Proceedings of the National Academy of Sciences.