Current Awareness and Evaluative Data in Academic Digital Libraries[*]

A proposal submitted to the OCLC/ALISE Library & Information Science Research Grant Program by

Thomas Krichel
Palmer School of Library and Information Science
http://openlib.org/home/krichel krichel@openlib.org

Submitting organization: Long Island University (LIU)
700 Northern Boulevard
Brookville, New York 11548-1300
U.S.A.

School/College Official: Jeffery Kane, Vice President for Academic Affairs

Participants:

Thomas Krichel

and

Sergei I. Parinov
Institute of Economics and Industrial Engineering
Siberian Branch of the Russian Academy of Sciences (SB RAS)
17, Lavrentiev Prospect
630090 Novosibirsk Russia
http://rvles.ieie.nsc.ru/~parinov parinov@ieie.nsc.ru

Abstract:

This proposal aims to examine the interaction between current awareness and evaluative data in a large digital library. We want to evaluate both the service and the parameters of the quality and performance of the documents that the service deals with. The current awareness service yields a stream of documents that are comparable enough for an evaluation of individual documents to make sense. The project will develop conceptual foundations, software, and a running implementation using the RePEc digital library. The implementation will generate a continuous stream of digital library usage data. All software and data will be shared with the digital library community.

[*] We are grateful to Bernardo Batiz-Lazo, Christopher F. Baum, David Goodman, Sune Karlsson, Marco Novarese, Kitty Rockett and Christian Zimmermann for comments on an earlier version.

1. Introduction

There are many difficulties in digital library evaluation research. One of the most serious is the lack of data. Web logs are too poor, questionnaire-based methods are too expensive. Only recently a workshop around this problem was held at the ECDL, see Larsen (2003). The NSF and Delos are aware of the problem. They have sponsored a series of workshop on these problems. But they have produced no testbeds. To produce a good testbed, one needs a large collection of data with heavy and sophisticated usage. Many collections don't have the scale that it takes, and if they do have the scale, they are not freely available; thus the copyright holders of the collection may impose restrictions on the collection and distribution of usage data.

As far as academic digital libraries are concerned, there are only three major collections that are basically freely available. There are arXiv, CiteSeer and RePEc. These collections differ in many important ways but share the fact that they have been built and continue to be maintained by dedicated individuals who want to provide a service to a target community over a long period of time. Likewise, longevity is crucial in the compilation of evaluation data. We need a long set of data to study. Therefore we need to run a technical infrastructure that will generate data over a long period of time, which we can then examine. We estimate that it will take several years to compile a useful dataset. This project, which will have to be carried out in a shorter period of time, therefore emphasis the setting up of an infrastructure that will generate data. The dataset will be made publicly available for all researchers to use. In addition, the project will do useful conceptual work and provide software in the process.

2. Background work

The participants have been working together on the RePEc digital library for many years; see references at the end of the proposal. RePEc dates back to the NetEc project founded by Thomas Krichel in 1993 and still operating today at http://netec.wustl.edu. The NetEc project is a

collection of services in the area of academic economics. Its largest component are bibliographic descriptions with full text links to working papers. In 1997, this work seeded the RePEc project. RePEc is digital library for economics. It contains descriptions of documents, collections of documents, researchers and research institutions. It is innovative in two ways.

The first important innovation by RePEc is that it does not have an official user service. Instead, it is a shared dataset that is contributed to by many participants and used by many user services. Currently, there are over 300 archives contributing data to RePEc. RePEc archives are based with academic economic departments, independent research centers, central banks, academic publishers and multinational administrations such as the OECD. There are around 10 different user services to RePEc. They are listed on the RePEc home page at http://repec.org. This separation between user data providers and service providers, as pioneered by RePEc, has been the model for the Open Archives Initiative. Thomas Krichel has been involved in setting up the initiative from the first meeting and served as a member of the technical committee.

The second important innovation is that RePEc is a relational database. Researchers and research institutes are described in access control records. The access control record for institutions are centrally maintained by a volunteer, and the records for individual researchers are maintained by the researchers themselves. RePEc is well on its way to truly becoming a community database for the research community in economics.

This application is concerned with developing further a user service known as “NEP: New Economics Papers”, see http://nep.repec.org. Thomas Krichel founded NEP in 1998. It provides current awareness services for RePEc. Every week, a volunteer with the help of special software compiles data about new working papers in RePEc. These data form the issue of a general report on all new papers. The report on all new papers is then circulated to a group of subject specific editors. These filter the general reports into subject-specific reports. At the time of writing, there are close to sixty such reports. Over the time of it's live, NEP has made over 30,000 announcements. Over 10,000 users are subscribed to at least one report. The history and

operations of the NEP service are reported in Barrueco Cruz, Krichel and Trinidad (2003), and Chu and Krichel (2003).

NEP is a simple, yet highly innovative service. First, it introduces a "push", rather than "pull" business model into the digital library. The library comes to the user, instead of relying that the user comes to it. Second, the current awareness business of NEP is shielded from the competition of Internet search engines. At a time when such search engines gain in popularity, the digital library community needs to find innovative business that can not be performed by search engines. Third, NEP breaks down the barrier between users and contributors. Fourth, the NEP service introduces an evaluation of the RePEc contents. This evaluation does not seem to be a vertical one, in the sense that it would rank papers or filter "quality" papers. Rather it is a horizontal one, much like the subject classification schemes that have been in use for long time. However, studies on NEP have shown that not all new papers that come into RePEc are announced. A significant proportion, up to around 30%, do not appear in any NEP report. Thus, there is a vertical evaluation as well. Though it is by no means as rigorous as traditional peer review, it complements such review marvelously well because it is cheaper and quicker. We will come back to this point later.

3. Evaluation of service and evaluation of documents

As we have pointed out in the abstract, this proposal aims to do two types of evaluations at once. First we wish to evaluate the NEP service. Here, we will examine the service as a whole as well as its individual reports. We need data on when the issues of each report come out. This alone is not trivial. We need subscriber data for reports at many points in time. We need precise download data for full texts. Currently the data generated by the running NEP service has a number of technical problems. We need more precise and more comprehensive data. Such data will allow us to build a battery of statistics to show report editors how well they are doing. The most important, and most challenging component of the service evaluation will be to build a system that will trace

the download behavior of each individual report recipient. To this end, we need to enhance the Mailman email list software that we are currently using for report issue delivery. The software, written in python, is open source; therefore it is possible to enhance it. The software currently sends out the same email to all subscribers. We need to personalize every email. Technically this will be handled by filtering every mail issued by Mailman before it is handed to the mailing software. A "ticket" will be appended to each URL that goes to a full-text papers download. The ticket will be an encrypted string, which, when decrypted with an appropriate key, will yield the NEP report, the issue data of the report and the email of the subscriber. The key to decrypt the message will be kept secret. We think that a secret key algorithm will be sufficient for our purpose. At fixed intervals in time, the web server that houses the downloading service will examine request logs, decode the tickets and establish which email address they are coming from. The project will then report, in a suitable XML format to be defined, anonymized records of downloads, where the email addresses of recipients are replaced by cryptic keys. Note that the ticket will not be used as an access restriction tool. Any user who wishes to download a paper without a ticket may do so. But such a download will be reported as an “anonymous” download. It will not count towards the evaluation of the paper that is being downloaded.

This idea leads straight to the second type of evaluation. We want to use NEP to combine the idea of current awareness with the idea of evaluative data for documents announced in the current awareness service. This requires further explanations. Within academic work, peer review provides the classic approach to evaluating research documents. From an abstract point of view, a peer review act consists of establishing whether a certain document can be a part of a group of documents. That group of documents may be a scholarly journal, or the papers presented at an academic conference; let us refer to them as peer review outlets to keep a general term. Usually papers within the same outlet have the same status. In addition, there is a perceived hierarchy of outlets for every academic discipline. What we want to examine with this proposal is a situation where papers in each channel are ranked and the channels themselves are ranked, too. The

ranking of the channels will be done as part of the service evaluation. The ranking of the papers within each channels will be done through the number of downloads. Initially, we will set up a system that counts the downloads of each paper from every issue of every report. We will publish these statistics through all RePEc user services. We expect that, as a result, some authors and their agents (family, friends, students) will seek to boost the positions of their paper by downloading them. This may appear silly behavior, and we are not sure to which extent it will occur. From private conversations with Simeon M. Warner, the system administrator of arXiv, we know that arXiv does not publish any usage figures precisely because they fear that these measures will be abused. Since our measures work on the disaggregated scale of the single report issue, they can be abused more easily. We would like to measure this effect for one year. After one year, we will foil the attempts of authors to increase their own rankings by introducing the ticketing system. Only downloads with a valid ticket will be counted. Each ticket will be given one "vote" or the most popular paper in the issue.[1]

While the evaluation of papers though counts of ticketed downloads does not have the same quality as real peer review, it will be very interesting to see what predictive power a high NEP download ranking of a paper has on the subsequent performance of a paper, as judged by its publication in a highly ranked outlet or by it citation count. Whatever the outcome, since it takes on average four years for a paper in economics to be published, and it takes many years after for the paper to become cited, it should come as no surprise if the profession will take a keen interest in an evaluative method that will deliver results comparatively fast.

4. Budget and administrative details

The project will run over a four years. The majority of the funds will be spent in Russia, where each dollar spent on computing support buys a multiple of what it buys in the United States. This

[1] We agree that ticketing is not a watertight system. But in practice, it will be very cumbersome for authors to circumvent ticketing, because they have no way of generating valid tickets by themselves.

accounts for the low cost of the project.

Travel: $2,800

Each year, the participants will meet in Novosibirsk for three months, roughly between 1 June and 30 August. Thomas Krichel will claim expenses for his travel and visa up to $700 per annum. He will pay from his own monies should the cost exceed $700.

Sub-award to SB RAS: $9,200

These funds will support a programmer at $1,800 per year. $500 per year is earmarked to offset the overhead expenses of SB RAS. They will hire the programmer. They will provide an office for him and Thomas Krichel. They will provide electricity and Internet access. LIU will arrange for the funds to be transferred to Russia. The payments will be made in May of each year. LIU has previous experience transferring to Belarus in support of Thomas Krichel's research.

Other expenses: $2,000

Since the project will be running for four years, it is prudent to account for additional expenses and cost increases. Therefore, we earmark as additional expenses a sum of $200 in the first year, growing by $200 every year, i.e. $200+$400+$600+$800=$2,000.

These additional funds also cover the administrative cost occurred by LIU. Any unclaimed or unused funds, one year after the end of the project, will become the property of LIU. They will be used to cover any additional expenses that may be incurred on the project.

Total request: $14,000

For the case of any dispute over the funds, the participants and LIU appoint Ivan V. Kurmanov, of Minsk, Belarus, as a mediator. All will adhere to his judgments.

5. Institutional support

Thomas Krichel estimates that he will spend 20% of his research time in the summer on this project. Roughly, at a wage with benefits of $60,000, that is around $3,000 a year or $12,000 for all four years. Other institutional support comes from the organizations that currently sponsor

RePEc services. The economics department at Washington University in St. Louis sponsors the bandwidth, CPU time and disk space for the current NEP service. They will also sponsor this service in the future. The hardware for the analysis of the log files and the Institute will sponsor the provision of a full-text cache for Economic Research at Hitotsubashi University, Tokyo. The hardware is already in place, a true-64 bit Hitachi machine with 8 processors working in parallel, each having access to 1 Gigabyte of memory, with a total disk space of 300 Gigabytes. It is difficult to quantify the precise monetary value of these contributions over four years. $2,000 is probably a reasonable figure. Thus, total institutional support is $14,000.

6. Time plan

The development work is carried out in the summer only. During all other times the participants will maintain the NEP system. But they will make no enhancements to its functionality. There is no sense to proceedings in any other way because we need the time during the year to evaluate the steps carried out in the summer.

Summer 2004

- complete remodeling of the back-end of NEP, including web interface for report generation
- historical cleanup of data and publication on non-ticketed usage logs[2]
- publication of non-ticketed download data

Summer 2005

- design of the ticketing system itself
- implementation of ticketing system, but no announcement to user community

Summer 2006

- setup of ticketing log processing system
- publication of ticketed statistics
- marketing of ticketed statistics to the economics community

Summer 2007

- publication of complete datasets for the digital library community, with a comprehensive

[2]Thomas Krichel has already done part of this work.

documentation

Each year, the participants will provide an annual report. Therefore this project, rather than being reported upon once, will generate four reports. The last report will be the longest because it will detail the data format and access.

The project is usual because of the length of time that it requires and the comparatively small amount of resources. The project participants are amply qualified to conduct such a project. From private conversation with Bill Arms, we understand that Thomas Krichel has established a reputation for getting a lot done with little resources. Sergei Parinov has been involved in a range of funded projects. Most recently, he has conducted work funded by the European Commission and by the Ford Foundation.

7. Conclusions

This project will conduct an experimental analysis to evaluate, on a large scale, documents through their usage in a digital library. The focus is not on comparing every document with every other document—that would make no sense—but rather on finding out comparable documents both by their subject coverage (as in the NEP report) as well as the time of appearance within the NEP report. Therefore there is a crucial dependence of the evaluation process on the current awareness filtering. This will be the main focus of the project.

The project will generate a very large datasets on the interaction between users of reports, documents in the reports, downloads and above all, two dimensions of time, the time of report issue and the time between the issue and the observation of its effects. We, the participants, can not hope to make justice to this data on our own. It simply is too vast. Instead it will be disseminated to others. This befits our way of working. We are very enthusiastic about the Internet as a tool to generate and disseminate freely available highly quality information. We hope that this application is evidence of that enthusiasm.

Finally, we hope that we will convince others that current awareness is an important service for

digital libraries. At the time of Google and Amazon, libraries need to seek new ways to promote their usefulness. Current awareness services will be part of a winning formula.

References

Barrueco Cruz, José Manuel, Thomas Krichel and Jeremiah C. Trinidad (2003) "Current awareness in a Large Digital Library" presented at the 2003 conference on users in digital libraries, in Espoo, Finland, September 8 to 9, see http://openlib.org/home/krichel/papers/espoo.pdf.

Chug, Heting, and Thomas Krichel (2003) "Current awareness service of the RePEc digital library: Progress, performance and potentials", to be presented at the Chinese Academy of Sciences Symposium on Libraries' Sustainable Development & Innovation, in Beijing, China, December 20 to 23

Larsen, Ronald (2003) "Digital Library Evaluation---Metrics, Testbeds and Processes", http://www.sis.pitt.edu/~ecdl2003/overview.htm.

Krichel, Thomas and Sergei. I. Parinov (2002), "The RePEc database and its Russian partner Socionet", Russian Digital Libraries Journal vol. 5, no. 2.

Krichel, Thomas, David Levine and Sergei I. Parinov (1999), "An Active Information Robot as a Network Agent for Researchers (illustrated through the RePEc/RuPEc online resources in Economics)", First Russian National Conference on "Digital Libraries: Advanced Methods and Technologies, Digital Collections", Saint-Petersburg, October 19-21.

Krichel, Thomas, Victor M. Lyapunov and Sergei I. Parinov (1999) "Online Scholarly Information for Economics: The RePEc database and the RuPEc web portal", Russian-British Digital Libraries Workshop, Moscow, June 16-17