Human mediated current awareness in a large digital library

José Manuel Barrueco
Jose.Barrueco@uv.es
Universitat de Valencia. Biblioteca de Ciencias Sociales
Campus del Tarongers, s/n 46071 Valencia

Thomas Krichel
krichel@openlib.org
Palmer School. Long Island University
720, Northern Boulevard
Brookville NY 11548-1300, U.S.A.

Jeremiah C. Trinidad
jadedlime@hotmail.com
Palmer School. Long Island University
720, Northern Boulevard
Brookville NY 11548-1300, U.S.A.

Abstract This paper presents and analyses NEP, the current awareness service of the RePEc digital library. NEP is a human-mediated service. New items arriving in RePEc are examined by editors of subject-specific reports. This paper introduces NEP from a conceptual point of view and communicates how NEP fits into the evolving world of digital libraries. We then present summary statistics for the performance of NEP. We pay particular attention to the coverage ratio, and the redundancy of reports. Suggestions for improving the performance of NEP are discussed.

1 Introduction

We are currently witnessing the stone age of digital libraries. This is a time when, for the first time in the history of mankind, collections of purely digital documents are here to rival, if not overtake, the printed library collections as far as size of data and accessibility is concerned. Seen from this angle, it will come as no surprise that the operation of digital libraries, as commonly understood, closely resembles the business of physical libraries. Typically, the digital library is a structured collection of documents made available through an interface of its own, just like the physical library is an organized collection of printed documents that is made available through its own interface, i.e. the library building, shelves, staff etc. This early analogy of the digital library and physical library also implies a distinction between the providers of a digital library, and its users. We can sum up these parallels between physical and digital libraries under the heading of the "legacy model" for digital libraries.

Some recent developments have started to push digital libraries out of the legacy model. Some digital libraries are collections of data that are used through several interfaces operating independently and simultaneously. A classic example is the Open Directory Project, see `http://www.dmoz.org`. Its RDF-like descriptions of web sites are created by volunteers. They are assembled in a centralized administrative structure maintained by Netscape Communications Corporation. They are then given to search engine to set up subject-tree architectures. These run in parallel with the traditional search interfaces that web search engines provide. More close to the subject matter of traditional libraries, we have another example in the RePEc collection of digital data about economics, see `http://www.repec.org`. One important feature of RePEc is that the collection is both composed and used in a decentralized fashion. That is, there are hundreds of contributing archives, who furnish data about documents, and possibly the documents themselves. They contribute to a collection which has sufficient structure to function like a conventional abstracting and indexing database. This database is then used in many services. This basic *modus operandi* from the RePEc database has more recently been extended and more formally standardized in the Open Archive Initiative's protocol for Public Metadata Harvesting since the year 2000. This protocol has received widespread attention. This is a clear affirmation that the business model pioneered by RePEc in 1997 is an interesting one.

In this paper, we consider another pioneering piece of work coming out of the RePEc community. It is the NEP current awareness service for new additions to RePEc. This is a human-mediated current awareness service. The idea is that new additions to RePEc are circulated to a group of editors. All editors specialize in a certain subject. These then filter the new entries manually into subject specific reports. Issues of these reports are circulated via dedicated email lists. These lists deal with announcements of papers only, they are not discussion lists.

NEP is technically quite trivial. But it is a pioneering digital library initiative. It goes beyond the legacy model of digital libraries. First it breaks down the separation between users and providers. Some users, the editors, have decided to make a log of their work with the collection public and share it with others. Second, the NEP service is not a pure service provider. It adds information to the RePEc collection. In this sense it goes beyond the legacy model.

In Section 2 we describe NEP more in some detail. Section 3 presents a simple assessment of the operations of NEP to date. There we many focus on the completeness of coverage. Section 4 examines the opposite problem to completeness of coverage, i.e. redundancy between reports. Section 5 discusses alternative approaches to improve NEP. The final section concludes the paper.

2 The basic idea of NEP

The origin of NEP is an idea by Thomas Krichel to create a human-powered current awareness list for the RePEc digital library. The name NEP was coined by Sune Karlsson. It stands for New Economics Papers. The service has a homepage at `http://nep.repec.org`. The basic idea is as follows.

There is a series of reports on new additions to RePEc. Each report is called a NEP report. Each report contains the new additions to RePEc that pertain to a certain subject, according to the judgment of a person called the report editor. Each report takes the form of a serial, i.e. it has a number of issues. Each issue is dated at the time when it appears. Report editors are free to issue issues of the reports as and when they see fit. Each issue is circulated as an email to a list of recipients, using mailing list software . Reports are identified by a handle that obeys to the case-insensitive Perl regular expression `nep-[a-z]{3}`. A special code nep–all is reserved for a list of all the papers that have arrived. Users can subscribe to nep–all, like they subscribe to any report. But nep–all is not a NEP report because it has not been edited to contain only papers of a certain subject. It contains all the papers that are available to the editors.

The York protocol defines the role of a general editor. This is a person who is in overall charge of the substantive aspects of the service. The general editor accomplishes several important functions. First, (s)he hires editor for the reports and makes sure that they are included in the nep-editors mailing list. Usually, the editors are PhD students or junior university faculty. Each editor is responsible for one or more subject areas. The subject area usually corresponds to the editor's research interests, though extensive subject expertise is not required. Nowadays, the general editor examines CVs of candidates for editorship. But there is no formal process of editor selection. Second, s(he) runs the special email lists nep–ann and nep–all. These are mailing lists, not reports. The first contains general announcements of the service. The second is a report-formatted data about all new papers. The items that flow into NEP are those that appear on the nep–all list. Finally, the most important overall task of the general editor is to monitor service quality. Clearly with close to sixty individual reports this is a daunting task. How to come technology can be called in to help is an important issue that we will come back to later.

The technical implementation of NEP has largely been the work accomplishment of José Manuel Barrueco Cruz. Each week, a script calculates the most recent additions for the working papers in the RePEc database[1]. Then it prepares a proposed report issue. This has the format of an actual issue, i.e. it contains the name of the report, the name of the editor and the date of this issue. It contains bibliographic information in two sections. First, a header has titles and authors only. The header section is followed by a body section with the bibliographic information as complete as the RePEc dataset affords, that is, possibly with abstracts and with URLs to full texts.

The proposed issues are circulated by email to a group of editors. Each issue arrives at the editor's inbox with all new papers that have been added to RePEc. The editor then weeds through the report to eliminate all the papers that do not belong to the subject matter of the report. (S)he has to do this both on the summary data and the full data section. A web interface for the composition of reports is also available. For security reasons, it is not publicly advertised. At this time, we are not aware of how many editors use the web interface versus how many massage the proposed issue in a text editor.

3 Overall empirical assessment of NEP

In this section we are doing some simple overall performance evaluation tests on the historic NEP data. Figure 1 shows the history of creation or reports over time. The birthday of the list can be calculated in two ways. First we can use the minimum of the issue dates all issue. Second we can use the minimum of the mail dates of all issues. We calculated both, and then took the maximum of both numbers. From our own experience this seems a reasonable empirical approach, though, clearly we do not show a coherent set of birthdays. The history seems discontinuous. More than half the reports were created in the first year. After that phase, in a period of time between April 1999 and August 2001 virtually no report was created. Then several lists appear to be created at the same. In recent days, no new report has been created.

In Figure 2, we look at the input into NEP. From the graph, there is an impressive increase in the frequency and size of the inflow to NEP. Individual nep–all sizes are subject to important fluctuations, however. There are periods where there has been no report for several weeks. These come as a result of technical difficulties. But even it times of a relatively regular sequence of nep- all issues, there seem to be a high volatility of the size. This is quite problematic, but there is little that NEP itself can do about it. At some times, the number of papers in nep–all reaches the dizzy heights of over 500 papers. In addition, we witness a rapid succession of nep–all issues in recent times. Wading through these piles of documents is by no means a simple task for the editors.

Figure 3 shows the coverage ratio of NEP through time. By that we mean the number of papers that receive at least one announcement in a report, divided by the total number of papers in nep- all, for each issue of nep–all. We should expect that the coverage ratio increases, as there has been an expansion in the number of lists. But it appears that the coverage ratio is static at best. The number of reports increases over time; but there are more and more papers to be dealt with. In this situation, editors are either overwhelmed and do not perform their job properly, or they become more choosy. Both effects decrease the observed coverage ratio. Note that it should not come as a surprise that the coverage issue falls off at the end of the time period. At that time, the very latest nep–all issues have not yet been filtered into reports.

We can see the impact of the size of nep–all on the coverage quite clearly by graphing size of nep–all and coverage ratio in a cross-sectional rather then longitudinal plot. Figure 4 shows this graph. When nep–all is very small, the fact that a single paper is missing has an important impact on the ratio. Despite this artifact of small numbers, there appears a clear negative relationship between nep–all size and coverage ratio. On this graph, there appear to be a couple of out-

[1] The RePEc database holds both working paper and article data. Working paper data describe papers that report recent research findings prior to formal publication. Article data concern peer reviewed papers. NEP, at moment only looks at working paper data only. This was a deliberate decision at the time when NEP was set up. The main reason is that the peer review process takes very long in economics. Delays for three years, not counting resubmissions, are common, and with resubmission, it can take five years for a paper to get published. Thus articles are not exactly new papers. In fact research active economists, especially at the top end of the profession, work with working papers, or even drafts that are circulated through private channels

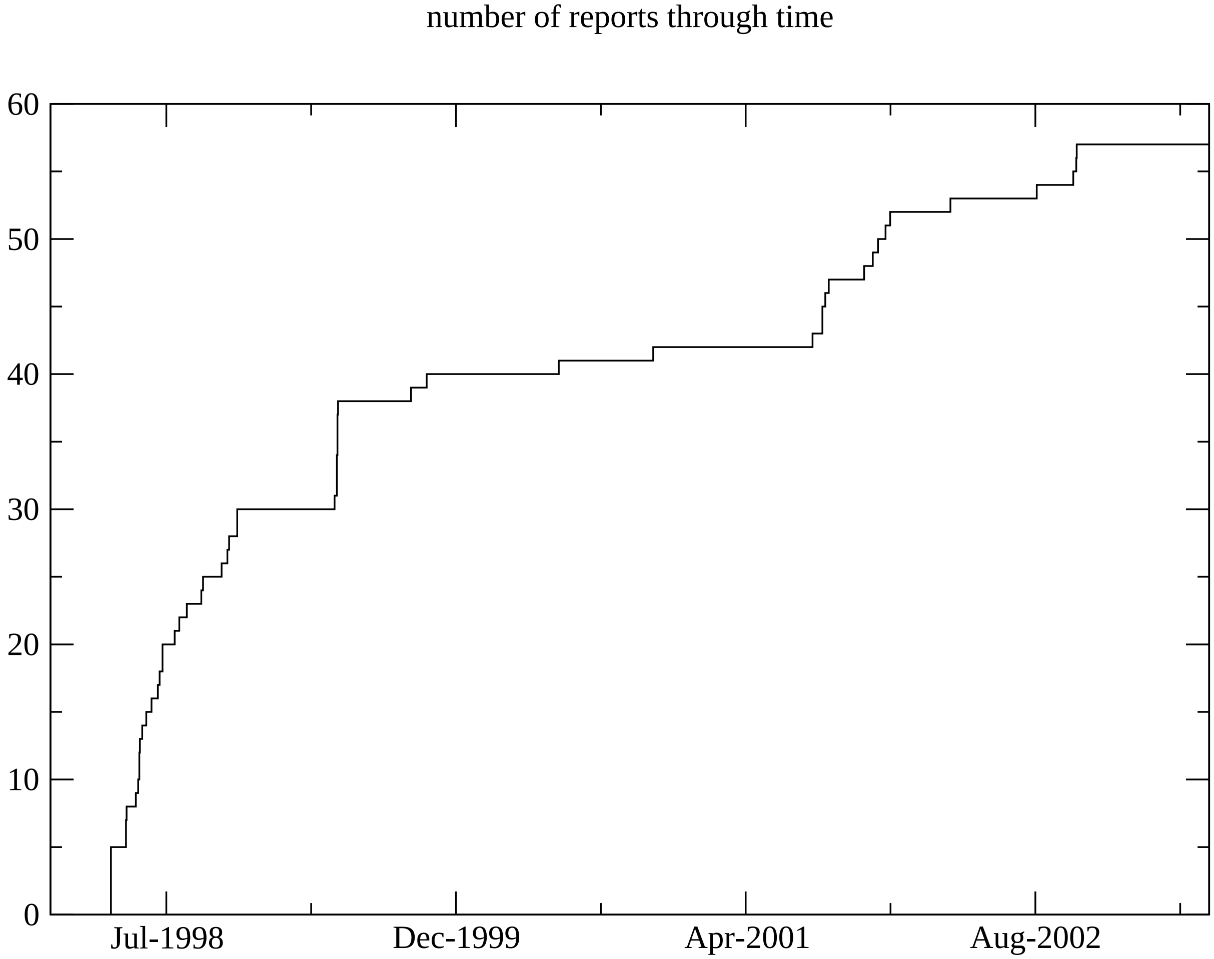

Figure 1: The number of NEP reports over time

liers where we reach a high coverage ratio, despite a large nep–all size. They would need to be investigated further. In addition, we could complete the picture by bringing in other factors, in particular the number of lists, in a full regression analysis, and to look at statistical techniques that would allow us to capture the stock effect of a number of large inflows that are coming one after the other.

An alternative way to grasp the coverage ratio of NEP is to look at it from the perspective of individual papers. Each paper may receive zero, one, two etc, announcements. In our Figure 3, we show the potential number of announcements, and the number of papers that receive that many announcements. It is interesting to note that despite the impressive array of reports, the number of announcements is not a multiple of the number of papers. It is also interesting to see that there do not seem to be many papers that are propagated through multiple lists.

4 Measuring the redundancy of reports

The development of NEP was not an exercise of careful planning to achieve full coverage from the outset. Instead, reports have opened as founding editors volunteered to edit them. When the funding editor retired, a replacement was readily available from the list membership. In this section we are looking at an objective measure for overlap between reports. This is what we refer to as redundancy.

The basic idea is a simple one. An announcement of a paper p on a report r is redundant to the extent that there is a user of report r, who subscribes to another report r where paper p is announced too. To fix ideas, imagine, as an example, two reports that are identical in the sense that they have the same papers announced in them. Provided that they do not have an overlap in readership, they are not at all redundant. Or, to take another extreme case, consider two re-

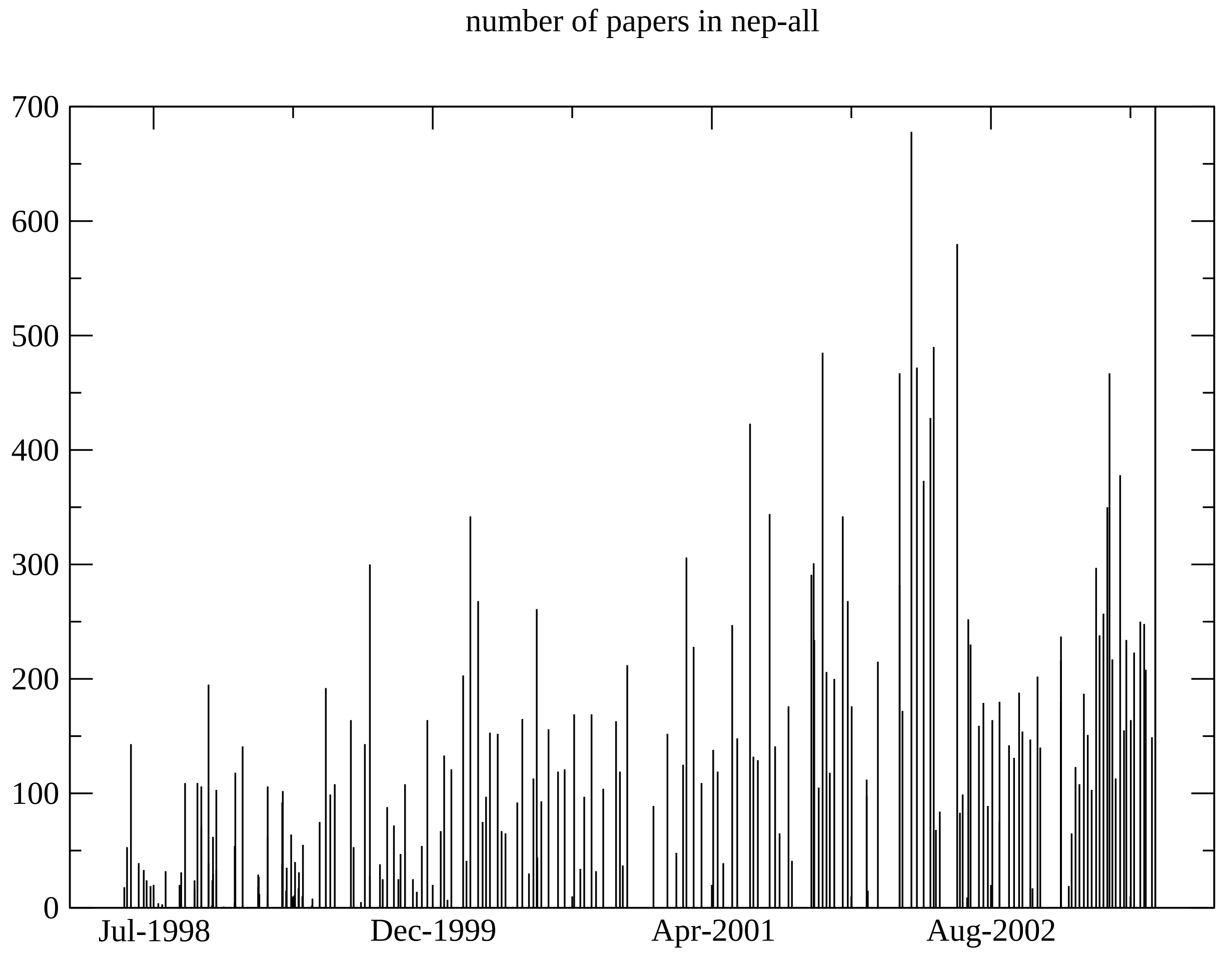

Figure 2: The size of nep–all

ports with the same users. They are not at all redundant provided that they announce different papers all the time. Only the occurrence of common users and common papers make reports redundant. Thus, redundancy between reports r and r' is the fraction of papers of report r that also appear in r' multiplied by the fraction of users of report r who also read report r'. Since the redundancy between two reports is a multiplication between two percentages, it is a small number. The redundancy of a report is the sum of the redundancy between itself and the all the other reports.Thus, while the redundancy between two reports is a small number, the total number of redundancies between a report and all the other 57 reports ends up adding up.

In Table 5, we list report identifiers in the first column, and usefulness in the second column. Usefulness is 100% minus the redundancy of the report expressed in percentage. We have ordered the list by usefulness in order to list the least redundant report first. The rest of the columns show the birthday of the report, the number announcements it has issued since birth, the number of subscribers, and the subject of the report. The main purpose of these additional numeric data is to show that there is no obvious way the usefulness of a report can be directly linked to its age or its size in terms of users or papers. Redundancy is an important feature of the reports at the bottom of the table. These will require the attention of the NEP management.

Two remarks are on order here. First the measurement uses subscriber data from 2003–06–01, but relies on data for the announcements of papers since report birth. To precisely measure redundancy we need to have data on which users receive precisely which announcements. This requires continuous time monitoring of the mailing list. We are not aware of how this can be done. While precise measuring is difficult, we could, in the future, do a better job than we have done here if we accumulate user data over many instances in time.

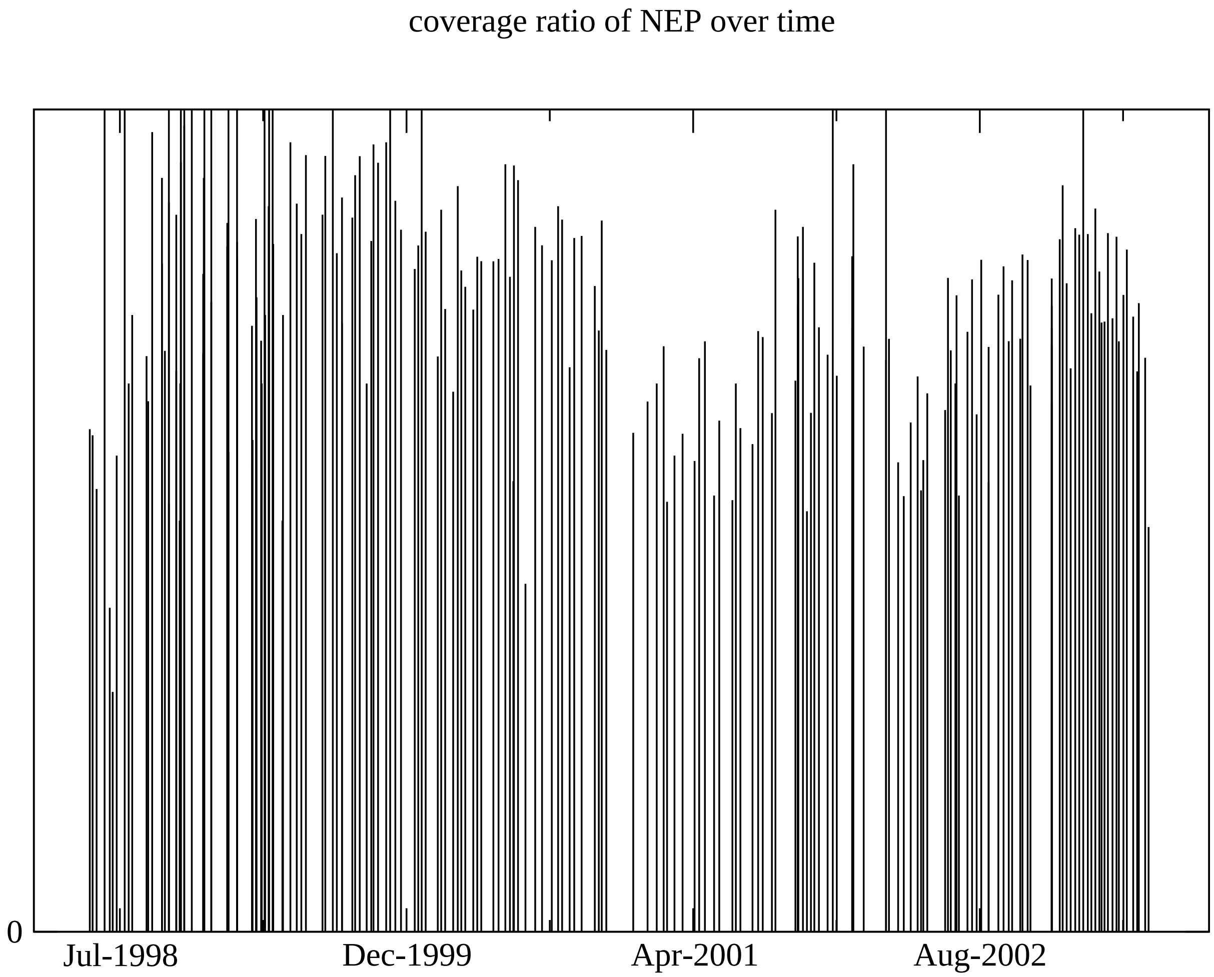

Figure 3: The coverage ratio, between 0 and 1

Second, the proof of the pudding is in the eating. Even if a list is redundant, it can still bring an important contribution because it can be a source of many full-text downloads of papers. In his paper, we have not looked at download data. This remains to be done.

5 Improving the operation of NEP

This review has focussed on the history of NEP and then examined in more detail two intuitive measure of the success of the system, the coverage ratio, and the redundancy. In looking at the coverage ratio, we have been mainly interested in the idea that we want to achieve comprehensive coverage. The idea of comprehensive coverage may not be appealing in a situation where the quality of submitted documents is doggy. But in the RePEc case, it is institutional archives that submit papers. Not every paper is a major scientific breakthrough, but if the only 70% reach one of the lists then we do have a serious problem of coverage.

There many ways to we can try to improve the coverage ratio. First, we need to lean on editors who are not doing a proper job. This involves calculating immediacy indicators. These are average delays between the email time of a paper in a report and the time of appearance of the same paper in nep–all. At the outset for the work on this paper, we wanted to report such figures in this paper; but the unreliable nature of the historical date data made it too difficult. Immediacy indicator research will have to be conducted in the future.

In order to cope with large inflows of papers, we can first think about a job-sharing protocol that would allow spreading the load of editing between different people. The York protocol explicitly introduced editorial teams but made no formal provision for job sharing. A second way to ease the workload on editors would be to either smooth out or restrict the number of items in nep–all. At the moment, we have all working papers flowing

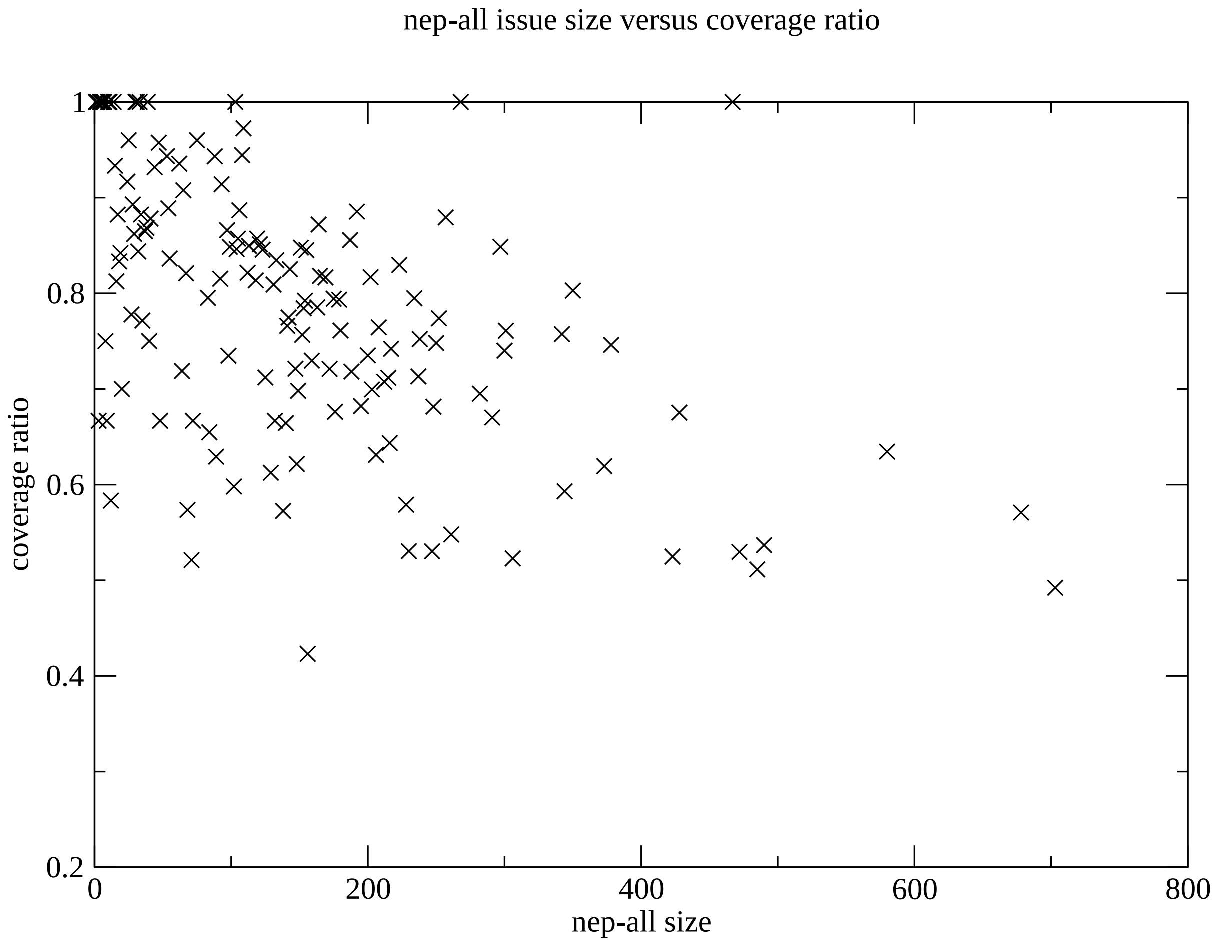

Figure 4: Size of nep–all versus coverage

in. We could restrict this only to working papers that are freely available online and for which we have secured the correctness of the full text URLs. Such a procedure has been proposed on the nep-editors list but no agreement could be reached on a way forward.

A third approach to reducing workload would be to introduce an editorial hierarchy. A hierarchical NEP would have first-tier editors who make decision on broad topics, and then leave it to second-tier editors to make decisions that would be communicated to final users only. While this idea has some intuitive appeal, it has many drawbacks. First, it would mean that the responsibility for reports is dissolved. Second-tier editors could blame first-tier editors for delays. Second, there is no good overall subject classification scheme to be used. Even if there were such a classification, implementing it now would mean revising the entire structure of NEP reports. This would damage the efforts of the best editors to build a brand name for their product. The best editors are the people we can the least afford to lose. Thus, while hierarchy is a good scheme to work on at the outset of a current awareness system, it is no use as a proposal for reform.

Therefore, with an unchanged lists structure, opening more lists could be a good idea. If there are many specialized lists, the editors could always take a narrow view of the subject, especially if they are aware that editors of surrounding reports may pick up a paper that they are not sure about. The only drawback is that from a users' point of view, someone who is new to NEP will have a harder time figuring out what lists to subscribe to.

Redundancy calculations are a good way to examine the structure of provision. Unfortunately it does not give us a glimpse for the gaps in the coverage. However, is it quite likely that highly specialized reports will be less redundant, and so will be reports that reach a special audience. Editors will have to be advised that if their contents is not very specific, they need to search for a special audience. They can do that by clever advertising. But from the analysis conducted here it seems that the creation of more specialised reports, without initial concern for overlap, seems to be a good way forward for NEP.

6 Conclusions

NEP is a simple, yet innovative effort. It pushes way beyond the legacy model of digital libraries. First, the users do not need to contact the library, instead the library comes to them, or, more specifically, to their email boxes. Second, NEP has "recent changes" mode of operation that can not be achieved through searching the web with a tool like Google. At a time when users are heavily turning to search engines to satisfy their information needs, NEP shows a distinctive advantage of human information organization over a vacuum cleaner approach. Third, NEP is another fine example of the RePEc ideal that with coordinated, decentralized volunteer efforts, great things can be achieved in the digital library field. Just examine of the service would be provided through a library of congress style classification apparatus. We just shudder at the thought of how much more costly this would be in both monetary terms and in time delays.

Finally, and most importantly NEP is an attempt to cross over the divide between users and providers of a digital library. One set of users, the NEP editors, have agreed to make the result of their usage of the digital library, the scanning of the lists of new additions, publicly available. The editors are therefore both users of the digital library as well as providers to it. While a lack of separation between users and providers are part of some Internet services, such as email lists, and personal web logs, it has hitherto received relatively little attention in the digital library literature. We think the digital library community should pay more attention to the potential of digital libraries to act as community tools. More generally, we firmly believe that the way forward for digital libraries lies more in the "animation" of the contents though user efforts, than in the aggregation of static contents in whatever sophisticated ways this can be done. In this paper, we have presented some of the trials and tribulations we had with a pioneering system. Implementers of similar system will be well advised to examine these issues before they are doing ahead with them.

7 Acknowledgements

The authors are grateful to the Joint Information Systems Committee (JISC) of the UK Higher Education Funding Councils for funding the Mailbase and JISCmail services that ran the NEP lists for most of its life. JISC also helped the WoPEc project, a RePEc user service that has continuously supported the development of RePEc. The authors are grateful to Heinrich Stamerjohans for advice with the graphics package "Grace", and to Heting Chu and Christian Zimmermann for helpful comments on an earlier version.

Id	**usefulness**	**birthday**	**#papers**	**#users**	**subject**
nep–spo	94	1998–07–20	24	1464	Sports and Economics
nep–ure	93	2002–10–24	256	139	Urban and Real Estate Economics
nep–com	92	2002–10–23	409	435	Industrial Competition
nep–ent	92	2001–08–16	894	317	Entrepreneurship
nep–lam	92	2001–08–16	314	616	Central and South America
nep–cul	91	2002–10–18	19	73	Cultural Economics
nep–pbe	90	1998–04–28	1151	1371	Public Economics
nep–hea	89	1998–04–27	702	274	Health Economics
nep–res	87	2001–11–06	99	239	Resource Economics
nep–lab	87	1999–04–22	2260	497	Labour Economics
nep–geo	86	2002–03–20	309	131	Economic Geography
nep–cbe	86	2002–08–16	188	128	Cognitive and Behavioural Economics
nep–his	85	1999–04–28	740	433	Economic History
nep–ltv	85	1998–09–04	741	861	Unemployment, Inequality and Poverty
nep–dev	84	1999–04–28	1368	477	Development
nep–dge	83	1998–06–24	929	476	Dynamic General Equilibrium
nep–edu	82	1999–04–27	182	1398	Education
nep–env	81	1998–08–10	535	452	Environmental Economics
nep–dcm	80	1998–07–28	330	313	Discrete Choice Models
nep–agr	80	1999–04–27	476	247	Agricultural Economics
nep–hpe	80	1999–09–01	333	238	History and Philosophy of Economics
nep–law	79	1999–04–28	572	247	Law and Economics
nep–eff	79	1998–06–01	175	416	Efficiency and Productivity
nep–net	79	1998–09–07	553	317	Network Economics
nep–sea	79	2001–08–22	241	72	South East Asia
nep–gth	78	1998–05–18	616	540	Game Theory
nep–eec	77	1998–07–20	1216	475	European Economics
nep–mic	76	1998–04–27	1697	472	Microeconomics
nep–reg	74	2000–05–13	246	276	Regulation
nep–ind	74	1999–04–26	1134	523	Industrial Organization
nep–pke	73	1998–06–21	1234	236	Post Keynesian Economics
nep–evo	73	1998–05–21	439	382	Evolutionary Economics
nep–acc	72	2001–08–11	131	72	Accounting
nep–mon	72	1998–10–19	1320	655	Monetary Economics
nep–tid	72	1998–05–21	798	427	Technology and Industry Dynamics
nep–ias	71	1998–11–05	365	144	Insurance Economics
nep–exp	71	1998–04–27	327	273	Experimental Economics
nep–ifn	71	1998–06–29	2004	602	International Finance
nep–tra	70	2001–11–28	225	119	Transition Economics
nep–mac	70	2001–11–15	932	309	Macroeconomics
nep–ene	69	1999–04–27	455	222	Energy Economics
nep–afr	67	2001–10–22	176	61	Africa
nep–ecm	66	1998–04–27	1264	889	Econometrics
nep–cmp	65	1998–10–09	337	368	Computational Economics
nep–fmk	64	1998–06–10	1178	821	Financial Markets
nep–cfn	63	1998–10–22	801	489	Corporate Finance
nep–mfd	63	2001–07–25	370	114	Microfinance and Financial Development
nep–pub	62	1998–05–20	1017	408	Public Finance
nep–cdm	60	1998–05–25	823	281	Collective Decision-Making
nep–fin	60	1999–04–22	1392	681	Finance
nep–cwa	57	2001–12–06	42	50	Central and Western Asia
nep–ino	57	1999–09–28	487	273	Innovation
nep–cba	54	2000–10–23	702	430	Central Banking
nep–pol	51	1998–04–28	401	350	Positive Political Economy
nep–ets	47	1998–04–27	1004	698	Econometric Time Series
nep–rmg	40	2002–11–26	545	80	Risk Management

Figure 5: The NEP lists ranked by usefulness