WoPEc usage in 1999AD*

very preliminary and incomplete

José Manuel Barrueco Cruz
Biblioteca de Ciències Socials “Gregori Maians”
Universitat de València
Campus dels Tarongers s/n
46071 València
Spain
jose.barrueco@uv.es
http://www.uv.es/~barrueco

Thomas Krichel
Department of Economics
University of Surrey
Stag Hill
Guildford GU2 5XH
United Kingdom
T.Krichel@surrey.ac.uk
http://openlib.org/home/krichel
RePEc:per:1965–06–05:thomas_krichel

Abstract

This paper talks about the usage of WoPEc using 1999 vintage log data.

This paper is available online at http://openlib.org/home/krichel/barcelog.html.

*The work discussed here has received financial support by the Joint Information Systems Committee of the UK Higher Education Funding Councils through its Electronic Library Programme

1 Introduction

The WoPEc service for electronic working papers in Economics is one of the oldest digital library for reseach papers in the world. It opened in April 1993 on a gopher server at Manchester Computing Centre. It published the first online research paper in Economics at that time. WoPEc was founded by Thomas Krichel. He had the vision of an academic self-help project that would be free for both users and contributors.

The WoPEc project has been three things at the same time.

- It was a repository for academic documents. This is the repository function.
- Second, it was site that collected metadata about online reseach papers the full text of which is based in other repositories. This is the (metadata) collection function.
- It was a site that users could interact with to search and download papers. This is the (user) interface service.

In 1994 José Manuel Barrueco Cruz joined Thomas Krichel. We as a two-man band dominated much of what happened at the WoPEc project in that period. Over the Internet, we worked together to offer repository, collection and interface services. After two years of additional volunteer effort, we bit successfully bid for funding from the Joint Information Systems Committee as part of phase 2 of their Electronic Libraries Programme (eLib). Now José Manuel Barrueco Cruz worked full-time on the project. Thomas Krichel had a part-time affiliation with the project. However, it was clear that it was not possible to scale the effort simultaneously in the three project areas. The scale we aimed for is reasonable comprehensive coverage of electronic working papers in Economics. Therefore a specialisation of effort was needed. There had to be demonopolisation of all three project activity areas. Other agent would need to be recruited

The WoPEc team decided to concentrate on the second aspect of the activity, i.e. the collection of material. To enable for a more decentralized collection, the team devised a "Guildford protocol at http://openlib.org/acmes/docu/GuilP.hmtl" that would allow to open the collection in two direction. First metadata providers could supply data to the collection by opening metadata archives. In addition the total collection would be available on public access computer systems for the simultaneous usage in a number of user services. This collection of data takes a life that is independent from WoPEc. The collection is known as RePEc. RePEc was founded in 1997. It now contains largest distributed library of free electronic research papers in the world. RePEc is a system that is open for both contribution and use. This twin openness of the RePEc collection is the crucial feature of RePEc. Krichel (2000), provides a more general description to the concepts that are realised in RePEc.

The WoPEc team has been the main driving force behind the the rise of RePEc. Conversely, we have spent much less time attending the development of the WoPEc user service. This has been a deliberate choice. We think that many digital libraries fail because they do not manage to assemble a critical mass of content. The history of the eLib programme is full of projects whose user interface is well designed and technically competent but where the contents is poor. We believe that there is a systematic bias within the library culture as a whole—and within digital libraries in particular—towards analysing the needs of the users rather than the need of the contributors.

As a consequence, we have tended to neglect our users. This paper is the first time we are actually taking a closer look at logs of user activity. Before we do that is it meaningful to look at the other RePEc user services, in order to set the scene for WoPEc. We do that in Section 2. We then introduce

the features of the WoPEc service in Section 3. In Section 4, we consider the logs of searches and page withdrawal. In Section 5, we look at the download logs for WoPEc. Finally Section 6 concludes.

2 The other RePEc user services

One of the main features of RePEc is that the collection is open for the creation of user service. That implies that there is not a single official user service. Instead a whole range of user services have been created. In this section we review these user services.

There are basically two groups of user services. The first group are primary library service, where the focus is on the access to academic resources. WoPEc belongs to this group of services. Here are other services that are part of that group.

- IDEAS at http://ideas.uqam.ca
  provides an Excite index of static html pages that represent all Paper, Article and Software templates.

- NEP: New Economics Papers at http://netec.mcc.ac.uk/NEP
  is set of reports on new additions of papers to RePEc. Each report is edited by subject specialists who receive information on all new additions and then filter out the papers that are relevant to the subject of the report. These subject specialists are PhD students and junior researchers. They work as volunteers. At the time of writing, more than 3,200 different email addresses that subscribe to at least one list.

- Tilburg University working papers & research memoranda at http://www.kub.nl/~dbi/demo mate/repref.htm
  This site also operates a Z39.50 server for all downloadable papers in RePEc is available at dbiref.kub.nl:9997. The name of the database is "repref". The attribute set is Bib-1, and the record syntax supported are USmarc, SUTRS, GRS-1 (only string tags, tag type 3).

- RuPEc at http://www.ieie.nsc.ru/RuPEc
  is a server in Russian. It offers search facilities to Russian users. Its maintainers also provide archival facilities for Russian contributors.

- INOMICS at http://www.inomics.com/query/search
  not only provides an index of RePEc data but also allows simultaneous searches in indexes of other web pages related to Economics.

A second group of services those that do not focus of the resources (i.e. documents) themselves. Instead they describe the creators of the resources.

- EDIRC at http://ideas.uqam.ca/EDIRC
  provides a web pages that represent the complete institutional information in RePEc. These are the data that RePEc holds to describe institutions that are active in Economics research. These comprise academic departments and some government institutions like of example central banks.

- HoPEc at http://netec.mcc.ac.uk/HoPEc.html
  provides a personal registration service for authors of documents in RePEc and allows to search for personal data.

Thus we have to see WoPEc within that framework of competing services. We should also recall that the WoPEc usage only reflects a part of total usage that the RePEc data gets.

3 WoPEc

There are a few features that are worth mentioning about. First, WoPEc is a part of the NetEc project. NetEc is the world's oldest portal for academic Economics. It started on a gopher server at Manchester computing in February 1993. NetEc comprises

- Information on printed working papers on BibEc,
- Data about electronic working papers on WoPEc,
- Code for Economics and Econometrics on CodEc,
- World Wide Web resources in Economics on WebEc,
- Jokes about economists and economics on JokEc.

Other projects that are associated with NetEc

- "Resources for Economists on the Internet"sponsored by the American Economic Association, editor Bill Goffe
- EDIRC("Economics Departments, Institutes and Research Centres in the World") by Christian Zimmermann

The idea behind NetEc was to combine the finest Internet services on one site. The overriding problem in the early days of the Internet usage for economists was amass contents. In these early days, contents was very limited. Thus assembling different types of contents to create a critical data mass was crucial.

WoPEc has the historic advantage of integration into a large portal structure. The disadvantage is that NetEc/BibEc/WoPEc/RePEc etc are confusing for the average user. From an insider's historical point for view, the structure appears logical. However for the average punter who uses NetEc for the first time is likely be quiet confused.

To add to the confusion NetEc has three sites. The original home of NetEc was at the National Services at Manchester Computing. There are two mirror site at Hitotsubashi University in Tokyo and at Washington University in St. Louis. The idea was that to give local users a more speedy access to the NetEc data. In 1994, when the mirrors were built, this was certainly a more important task than today. Today it appears possible to run everything on a central site. For anybody who lives in one of the countries that currently have a NetEc site we would still have a reasonable speed of access. It remains that having various NetEc sites in the developed countries would do little to improve access for users in countries where the domestic lines are thin. The historic justification for the mirrors has thus largely disappeared. However, it remains that having the mirror site adds to a prestige of NetEc as a group that operates on a world-wide.

The split between BibEc and WoPEc has a historical motivation. Whereas NetEc received bibliographic data for many thousands of printed paper right from the start, WoPEc started as an empty content shell, because there were no electronic papers available when WoPEc started. In an common

index for printed and electronic papers these electronic papers would not have been found. Therefore it was clear that two different indexes where needed because the WAIS search engine could not—at the time—search several indices at once. The idea to make each index a separate NetEc project was only natural.

To conclude, the WoPEc project is embedded in a complicated user service infrastructure (NetEc), as will as in a complicated data provision infrastructure. This does not help to gather and it hinders the interpretation of usage data.

4 Page access

4.1 Derobotification

The NetEc services have been on the web since 1994. They are widely known to the public—and to robots. There are 6602099 lines in the log. We found 7201 hosts accessing `robot.txt` We removed all the line accessed by all those hosts. We end up with 1498332 derobotified lines. We then made files per host, by making sure that we aggregate all host names to ip numbers. This involved many DNS lookup and we lost more time over that. We found 138482 hosts.

Clearly these hosts are of very varied nature. They include cache and proxy hosts and firewalls. They include machines used by individual users in their offices as well as machines in computing labs that are shared by many users. One way to approach the individual user is to look at access hits from a single host that are in rapid succession. We call such a sequence of hits a session. We will say that a hit from a host belongs to the same session as the previous hit from the same host is loss than two hours ago.

Since each machine can access the service several times, we defined that a new “session” would have started when two hours have passed since the last line. In our view, this concept of sessions corresponds to an interaction of a user with WoPEc. We then split the file into sessions. We found 253506 sessions.

4.2 Session analysis

How many sessions are there per host? From the table there is substantial evidence of repeat usage. Only 14% of all hosts have only accessed the system once. The average number of sessions per host is ????. The number of sessions per host is of course not equal to the number of users because of a host serving many physical users. The total number of users is larger than the number of hosts. Therefore repeat usage is overestimated in these statistics.

number of session	percentage of hosts
1	83.12
2	8.72
3	2.72
4	1.34
5	0.81
6	0.57
7	0.41
8	0.29
9	0.25
10	0.18
11	0.17
12	0.14
13	0.10
14	0.09
15+	1.10

average: 1.83

How large are sessions. Here we look at the number of lines. Now this seems interesting, because the numbers of hits per session is small. It appears that a lot of pages are cashed. There also seems to be a preference of even—as opposed to odd—numbers of hits. There is something in here that we are missing out on.

number of lines	percentage of sessions
1	8.99
2	48.83
3	6.91
4	13.63
5	0.40
6	6.21
7	0.20
8	3.31
9	0.60
10	1.91
11	0.08
12	1.78
13	0.05
14	1.00
15+	6.11

average: 5.39

Here comes the same statistics for hits of papers. Thus for each number of papers that may appear in a session we give the percentage of session that contain that number of page entries.

It is interesting to note that the profile that emerges from there is quite similar.

number of papers	percentage of sessions
1	9.28
2	53.11
3	7.36
4	13.25
5	0.30
6	5.77
7	0.16
8	2.89
9	0.43
10	1.62
11	0.04
12	1.31
13	0.04
14	0.80
15+	3.64

4.3 Search Engine access

SQL 103.576
WAIS 281.376
ROADS 208.942

We found this data by searching through the log to find URLs that point to the search script

`zgrep '/`*xxx*`' log/*1999* |cut -f7 -d' ' |grep` *xxx* `|grep -c ^`

where *xxx* was *sql* for SQL, *wais* for WAIS, *search.pl* for ROADS. This implies that we count the number of times that someone executed a search, i.e. that the url of the search engine appears in the web log. This does not tell us anything about how useful these searches were. For that we would need to find out what the number of items retrieved after each search where. This can be done if we have a referrer log built in, but that log is not fully available in the data that we considered.

For those entries where a referrer exists, we can count the appearance of paper page withdrawals per referrer log entry.

Note

5 The download statistics

Many of the papers in RePEc are not part of the RePEc dataset. The live on URLs that are outside RePEc. RePEc only links to these papers. Under normal circumstances, it would not be possible for WoPEc to gather data on the downloads of these papers. The act of download would be recorded in the log of the remote site, not in our log.

Since we are conscious of the importance of downloading data, we do not link directly to papers. Instead, we link to a CGI script `download`, whose only purpose is to record that the paper is being downloaded before the download takes place. Since the search engines do not follow CGI scripts, there is no need to remove robot-generated entries from the downloading information.

Number of downloads 219,150

To compute an average download per paper, we need the total number of paper that were held during 1999. The number of papers in the collection is changing constantly. The compute the average stock of papers held, we would need to know, for each number of paper that the collection held at any point

in time we need to know for how long it has been held. Since we do not have this data available, we can have a reasonable (probably to high) number for the average stock as 15,000. This implies that every single document was on average accessed ??? number of times.

Downloads by months:

Jan	16479
Feb	19369
Mar	24500
Apr	25040
May	22876
Jun	17787
Jul	6673
Aug	5649
Sep	9258
Oct	23756
Nov	24100
Dec	23663

It appears that that the number of downloads has a seasonally that follows roughly the time period when American student have to write essays. In particular there is a substantial drop in the downloads at the time of the summer. It should also be noted that the total amount downloads has an upward trend, but it is not very large. For the first three months the average number is ?????, whereas for the last three moths the average number of downloads is ?????. This is an increase albeit not a spectacular increase. Since the usage is only about 1/3 in the summer than what it is is the Winter, we estimate that about 50% of total usage is student usage.

Of course, behind the overall document downloads, there is a significant differences in usage across documents Here we have all documents downloaded more than 150 times

1	RePEc:fmg:fmgwpswp0010	649
2	RePEc:hhs:hastef0260	369
3	RePEc:wpa:wuwpfi9609004	407
4	RePEc:fem:femwpa1999	346
5	RePEc:fip:fedaery:1996:i:Jan:p:1-20	309
6	RePEc:wop:ucbrpf272	269
7	RePEc:wop:scfiab_001	239
8	RePEc:bon:bonsfa484	239
9	RePEc:fip:fedgfe1997-13	235
10	RePEc:wpa:wuwpot9807001	229
11	RePEc:wop:aleapa_004	219
12	RePEc:wpa:wuwpfi9402001	219
13	RePEc:boc:bocoec318	205
14	RePEc:wop:aarhec1996-7	203
15	RePEc:wuk:elecwp9608	188
16	RePEc:fip:fedmqry:1995:i:Fall:p:2-17	185
17	RePEc:dgr:kubcen199610	177
18	RePEc:fip:fedbwp96-7	174
19	RePEc:wop:cercwp9603	174
20	RePEc:wop:frbfes9619	174
21	RePEc:nbr:nberwo6344	167
22	RePEc:wop:frbfes9620	164
23	RePEc:wpa:wuwpma9807002	158
24	RePEc:wpa:wuwpfi9411001	154
25	RePEc:nbr:nberwo5129	151

On the other hand, many papers are downloaded only infrequently. Note that we can not compute an average number of downloads since the log contains no information about papers that have never been downloaded.

number downloads	percentage of papers
1	14.24
2	10.35
3	8.86
4	7.58
5	6.25
6	5.43
7	4.46
8	3.89
9	3.64
10	3.11
11	2.77
12	2.33
13	2.19
14	2.27
15+	22.63

It should first be noted—as one would expect—most of the documents are old. Presumably all of them are pre-1999. Clearly it in not likely that a freshly arrived paper makes it to the top of the

"download chart", in particular if the basis of its compilation is a year.

The second point is that some documents seem to have been withdrawn. Most of those that have been withdrawn come from the RePEc:wop archive. This archive gathers data from sites that do not yet have their own archive. When on own archive is being built, then the templates in RePEc:wop are being withdrawn. There is currently no relational mechanism between a template and its successors in other archives. Thus a line of dependency can not be traced here.

Finally, the most interesting thing is that all paper that have high downloads are in the broad field of finance. The reputation of the author seems to matter much less than the contents of the paper.

Here the list of all hosts that have downloaded more than 200 papers:

194.225.36.2	Teheran University	1077
144.32.128.3	pump1.york.ac.uk	332
195.142.236.254	test.tcmb.gov.tr	254
137.205.8.1	bluebell.csv.warwick.ac.uk	253
134.83.176.45	lumen.brunel.ac.uk	251
193.205.23.1	dns.sm.uni-bocconi.it	246
130.115.115.82	few-115-82.seor.few.eur.nl	216
131.251.0.11	ramoth.cf.ac.uk	216
195.129.1.132	gate.caboto.it	216
147.47.1.102	Korean Education Network	215
200.13.213.70	completely.unknown	212

This chart of the heaviest downloading hosts seems to be composed out of cache and firewall sites. The majority are at academic or academic related sites. To really put the usage of various institutions together, one would have to aggregate usage from various machines within an organisation.

The last table shows the lower end of the host access. It appears that 61% of all hosts that have downloaded a paper only ever downloaded one paper. It would be interesting to see how much that would make in terms of the total number of hosts.

number of downloads	percentage of host
1	61.10
2	15.43
3	7.11
4	4.08
5	2.61
6	1.89
7	1.34
8	0.96
9	0.75
10	0.59
11	0.47
12	0.36
13	0.30
14	0.26
15+	2.76

6 Conclusion

It is quite difficult to gather meaningful results out of log files alone.

References

Krichel, Thomas (2000). Working towards an Open Library for Economics: The RePEc project. presented at the “PEAK 2000 Conference: The Economics and Use of Digital Library Collections”, available at http://openlib.org/home/krichel/papers/myers.html.