Budapest document

Thomas Krichel and Ivan V. Kurmanov

2002–07–17

1 Status

This document is the Budapest document. The latest version may be found at http://openlib.org/home/krichel/budapest.html. It contains an introduction to the RePEc project, with a development plan for its author registration service HoPEc.

2 What is RePEc?

After the ArXiv for Physics and related disciplines, RePEc is the second largest discipline-based Free Online Scholarship initiative in the world. RePEc documents scholarship in economics and some related areas. RePEc pioneered the OAI business model that distinguishes between data providers and service providers. There are around 250 archives that contribute to RePEc. Around 10 user services have been built using that data. There are 200,000 documents indexed in RePEc, more than a third of which are freely available. The project's web site at http://repec.org has more figures. It is maintained by a small team of volunteers. It has received indirect funding in the past from the Joint Information System (`JISC`) of the UK Higher Education Funding Councils, in 1996 and the following years, through support to the WoPEc project. The total amount of support was GBP 129,000.

3 RePEc is more than just a digital library of documents.

A conventional library is a collection of documents plus a user interface to search it. RePEc is not like this at all. It is different in many aspects. One important difference is crucial. RePEc is not just a catalog of documents. In addition to document metadata we collect data about:

1. institutions involved
2. researchers involved

Institutions' data has been collected and is being maintained centrally. A member of our team, Christian Zimmermann, does that, by maintaining a project called EDIRC. People write to him with their institution's (updated) details and he fixes the records accordingly. There are more than 5,000 economics-research-related institutions in his dataset.

To collect data about researchers themselves, we can not proceed in the same way, because the amount of data that is involved. We use a more sophisticated method.

4 Collecting data about researchers

In 1999, we created a special online service called HoPEc. The name stands for something like "homepage papers in economics". Economists come to it and register themselves. So far more than 4,000 researchers registered. They are providing us with their contact details, affiliation data and research data.

Contact details are the person's email address, phone number, postal address etc.

Affiliation data is a list of organizations that the person claims to be affiliated with. More technically speaking, it is a list of references to institutions already described in RePEc. Researchers are searching in the institutions database for appropriate ones by name or geographical location.

By research data we mean a list of papers, articles or software components that are authored or co-authored by the person. To be more precise, it is a list of references to the document items in RePEc. During the registration process of a person, the system makes a search in the RePEc documents database for the items which have a variant of the spelling of the registrant's name among the authors. The person then chooses relevant items from those found. We call this process "claiming authorship". The list of claimed research items makes up the person's "research profile".

Each person's registration is confirmed through email before it steps into effect. Thus it requires a valid email address to be entered during the registration. Otherwise, the service is open for anyone to register.

Once a registration is completed—or upon a later update by the person—the information enters the online registration system, and more importantly, it enters the RePEc catalog to be used by other RePEc services.

You can view a sample personal profile page at

http://netec.mcc.ac.uk/adnetec-cgi-bin/gemini.cgi?submit=id&HANDLE=RePEc:per:1945-02-12:DAVID_FRIEDMAN

Data behind this person's profile is available as a data file at

ftp://netec.mcc.ac.uk/pub/RePEc/remo/per/pers/RePEc_per_1945-02-12_DAVID_FRIEDMAN.rdf

A page on EconPapers RePEc service representing the same personal profile:

http://econpapers.hhs.se/HoPEc/49575253485049506865867368957082736968776578.htm

Most importantly, we gather log data for all the person's paper:

http://logec.hhs.se/HoPEc/49575253485049506865867368957082736968776578.htm

This is used to build a list of top authors at RePEc. It is also used to send registered authors a monthly email about how will they are doing. Each time the emails are sent out, we get lots of registration updates. This is a sign how well we are doing.

For more information on the existing service see http://authors.repec.org.

5 Backstage: the software we have now

The HoPEc system is now running on the software made by Markus J.R.~Klink, who was paid from the JISC grant. HoPEc was his graduation project. Since then he has not been able to devote much time to it, as he left the project for a job in a commercial company in India of all places. The software is written in C++ programming language, which is not bad by itself. Alas the interface was not carefully designed. When Markus installed its first version for our own internal testing, it caused a barrage of criticism from the RePEc bigwigs. To satisfy the critics Markus had to change the software logic significantly, but he was forced to do it quick. As a result, the software became populated with ad-hoc solutions and design flaws.

As bugs and design flaws in the software came up we had to fix and solve them. At the end of year 2000, Ivan V. Kurmanov took over maintaining the software. He put a lot of effort into understanding and fixing and rewriting parts of the code written by Markus and that by most part was successful. Some new functionality were added, some problems solved. But the existing system structure and the code structure in general is still too bad to allow painless extensions and improvements.

Now, after almost three years of running the service, we need to move forward. Many new ways to make service more useful to the public came to our minds. Some of them we believe are really demanded by the economics community. While taking

as much lesson as possible from the experience, we see only way forward: A complete rewrite of HoPEc software.

6 Where do we want to go today?

We aim at providing our users with a simple-to-use, secure and full-featured on-line curriculum vitae service. The main components of a researcher's personal vitae is contact, research and affiliation information. To an extent these components are already supported by the system.

We need to give people a bit more place to leave their personal information, like info about their research fields, their birth date, their photo and so on. All that, of course, will be optional for users.

Probably more important, we need to make system safer and easier to use. For improving user-friendliness we will fully utilize the feedback collected from the current system's users. We'll try to minimize the number of required "clicks" and page reloads during the registration process and following updates.

After all, the service will be completely redesigned, moved to a different network location (URL), changed name from HoPEc to "RePEc Author Service" (RAS).

7 Automated research profile

Another important thing to do is simplifying research profile maintenance. We want to make keeping one's research profile in RAS up to date as easy as possible.

One obvious aspect of this is to allow people adding works which are not in RePEc to their profile. This is possible now, but it takes unreasonable time and effort. It has to be streamlined.

Another issue is that many busy economists (especially, well-known ones) would not bother to come to our site regularly to claim the recently appeared articles and papers. We think an automated claiming system will be able to cure that. The system will study every new document addition to RePEc, checking if any of it's authors name matches that of any of the registered researchers.

If there is a matching registered person, the system will either automatically add that to the person's research profile or will simply invite the person to the site to claim or reject the item.

To do all this is not a problem. To do it well is.

8 A general tool

We are ambitious about the RAS project. We believe it is an important step on our path to enhance free scientific communication on the Internet.

RePEc is still focused on economics discipline. That is the environment we all came from. But there are other initatives that could use the software. Already there are there is the parallel rclis (Research in Computing, Library and Information Science), pronounced "reckless" project, which uses the same technology to collect research metadata in library science and computer science. At the end of the day RAS could evolve into a general tool to maintain scientific online directories.

To make our developments useful not only to ourselves, we need to document them well and design them carefully. Good documentation means other people can take the software and use it. Careful design means that later changes, customizations and extensions are possible and don't take a rocket scientist.

Experience shows: both are not easy, both take a lot of effort and time.

9 Time is money. We need both.

We estimate the cost of the project to be USD 4.400. Most of that will go for software development and related work, e.g. writing documentation. The software will be released as open source. Our software development relies as much as possible on existing open-source tools. The cost of the RAS software development is directly related to it's specific nature. We are not aware of any similar open-source project, whose code we could take and adapt to our needs.

A small part of the money will go for visual design of the service's web interface.

Implementation of the project is expected to take approximately 6 month. Because time estimates are usually unreliable, we would reserve two more month, summing up to 8 month.

10 Budget

All work to be done by consultants in Minsk, Belarus. All figure are in US Dollars.

item	*cost per time*	*number of times*
programmer, basic wage	200	8
programmer, bonuses	200	3
graphical designer	200	1
Internet cost	50	8
software designer	200	1

Total cost: 4400.