ArchEc: an archive for RePEc, part 1

A funding application to the Banque de France Foundation

2019–01–09

This application comes on behalf of the RePEc community. It is for 3000 Euros to start to work on a comprehensive archive of all the working papers that RePEc has links for. We hope that future applications to the foundation may address the wider preservation of RePEc data. For now, our application has two parts. In the first part, we produce an outline of what problems the funding addresses, and what work it will do to address the problem. In the second part, it will address the formal funding criteria that the foundation has set out in its documentation.

Part 1.

1.1 Problem statement

The RePEc project is well known, but poorly marketed and poorly understood. At its heart, there is a collection of over 2000 contributing “RePEc archives”. They contribute metadata about the working papers or articles that they publish. This metadata includes descriptive information such as author names, titles, abstracts etc. Access to the full texts is supposed to be handled via links. Most of the time, these links refer to the working paper full-text file in PDF format, but in some cases they go to intermediate pages which refer to the full-text. Working papers from most providers are freely available, while most journal articles are behind a paywall. The problem with this system design is that an archive can disappear; its full texts can disappear; and when papers are revised, older versions are gone. Without a historical archive, the economics community lacks access to old working papers, which often contain much more detail on the underlying research than what will be published in a journal. Researchers in low-income countries who lack access to toll-gated journals have no access to the working paper if it becomes unavailable in working paper form, and is only accessible in toll-gated published journals. As an aside, note that working papers are specific to economics. There used to be a working papers culture in computer science. Computer scientists called them “techreports”. Computer science techreports disappeared in the 90s as researchers put their papers on the web without any coordinating activity. Thanks to the RePEc project, working papers are alive and well in economics. Other scientific disciplines rely on a centralized archive, ArXiV, which is managed with a sizeable budget. In contrast, working papers in economics are generally freely available through RePEc, linked to published articles to encourage proper attribution. In other disciplines, such as psychology and public health, free circulation of research papers precludes their subsequent publication. As economists are the only research community making widespread use of working papers, we cannot just simply imitate what other communities are doing.

1.2 Limitations

To fully preserve the RePEc data and full-text is a task of considerable technical and social challenge. It cannot be done in one year. This application aims to start with the low-hanging high-yielding fruit of preserving the full-text that we can find in PDF format. The application will not aim to make the stored files available immediately. Without having consulted RePEc archive maintainers, we assume they prefer the papers to be accessible on their servers. It is only when the papers on their servers disappear that we should make them available. When the work starts, this is equivalent to the so-called “pickled” condition discussed below.

1.3 Extent

The extent of the work is 100 hours of work by Thomas Krichel. In the next section we define a set of work described as the warranted work. We guarantee that the warranted work will be completed, regardless of the time commitment required. We hope to extend the work beyond this benchmark.

1.4 Tools and standards

We aim for formal web archiving using the WARC format. This is a standard for web archiving: see http://bibnum.bnf.fr/WARC/. We use the Python programming language. Code will be in a Github repository.

1.5 Work

1.5.1 Warranted work

We will extract URLs of full-text working papers from the historic data of the NEP: New Economics Papers project, available since 2005. This is where the bulk of working papers can be found. There are 394,525 such papers at our last count. Some of them may have changed handles. We will use Sune Karlsson's handle change data to take account of that. There will be one WARC file per paper in RePEc that we aim to preserve. We will download what we find at these URLs. These payloads will be archived in one WARC file per paper. If the URL goes to a PDF file, the PDF file will be assumed to be the full-text instance of the paper. For the full-text that can no longer be retrieved, we will attempt to use the CitEc storage where José Manuel Barrueco Cruz has downloaded papers for citation analysis. If we find such a file, we will say that the WARC is pickled. Assuming that a stored copy exists, we assume it to have been available at the time stamp when we accessed the stored CitEc file. We expect that dealing with this data will take the largest part of the work. The WARC will contain a notes that the payloads come from a stored copy, rather than a live URL retrieval. Pickled WARCs will be made available for public access immediately.

1.5.2 Extended work

If end points contain HTML files, we will parse the HTML, and follow all the URLs found in the page. If the resource contains a PDF, it will be stored and considered as a full-text for the paper. We will carry out a page count check and check whether the title string and/ or author names can be found in the converted PDF. We will extend the full-text download to all types of data in RePEc. Thomas will integrate the WARC store into the CitEc and NEP delivery workflows. Finally, if there is time left, Thomas will work on a conceptual paper to bring in metadata into the full-text WARC.

1.6 Accounting

Thomas will keep an up-to-the minute ledger of all the 100 hours. The time to keep the ledger does not count for the time spent on the project. In fact, the time on the project is purely for coding and maybe a bit of system administration. There will be a short final report. We can create a mailing list for people interested in the project.

1.7 Conclusions

Thomas Krichel initiated what was to become RePEc in 1993. RePEc itself has been around since 1997. It has stood the test of time well. Its unfunded nature makes it very resilient. It last received external financial support in 2002. Clearly, the project is running on its own but on some occasion it needs external support. We need to roll out better, more reliable full-text delivery. This is a chance for the Banque de France Foundation to become part of this trail-blazing piece of infrastructure. In economics, old research rarely loses significance. Future generations will be grateful for the insight into our contemporary economic thinking that the papers can provide.

Part 2.

In this part, we look at the formal conditions as outlined in the document at https://fondation.banque-france.fr/sites/default/files/media/2017/05/17/application\_form\_conference\_sponsorship.docx

L'organisme est une société savante ou un réseau à but non lucratif dont les membres sont des universitaires, chercheurs ou praticiens et dont l'objectif est le développement et la diffusion de la recherche dans les domaines monétaire, financier ou bancaire.

This application comes from RePEc. RePEc is not an organization. When it started in 1997, it had no formal decision-making structure. In 2010, Thomas Krichel created an organizational structure that made RePEc an independently run project of the Open Library Society (OLS). The OLS is a tiny 501(c)(3) charity registered in New York state. RePEc has a board appointed using its own rules. All the board does is to approve resolutions. The board’s rules and resolutions are available at http://gove rnance.repec.org/. The latest resolution is the unanimous approval of this funding application.

Le rapport annuel de l'institution fait apparaître sa taille, par catégorie de membres. Le bilan des publications dans les revues académiques internationales et des réunions de toute nature témoignent de sa contribution à la recherche scientifique et de la dissémination du savoir dans les cercles de politique économique.

The OLS does not have members. It has no substantial influence on RePEc. It produces no account of what RePEc does. The RePEc board members are at http://governance.repec.org/board.html. The board passes resolutions. The count of actual numbers and types of records in RePEc is made by Christopher F. Baum. He maintains the web site http://repec.org. RePEc does no economics research itself. It works on the dissemination of academic knowledge in the economics.

La Fondation est identifiée comme membre donateur dans le rapport annuel et comme lien sur le site Internet. Elle est reconnue par son logo.

Neither OLS nor RePEc produce an annual report. RePEc’s web site at http://repec.org will be updated immediately upon receipt of funding. It is in our best interest to be seen being supported by the foundation. We will be happy to make additional announcement on the site about work by the foundation by placing links to their site on request. In addition, we will allow the foundation to place free advertising through NEP report emails throughout the course of the funding period.

L'adhésion ouvre droit à certaines contreparties telles que des abonnements à une revue ou des inscriptions à un colloque.

We will be open to suggestions as to what these counterparts should be.

L'association est représentée dans le conseil d'administration ou le conseil scientifique de la Fondation, ou réciproquement. Elle désigne un représentant au titre de membre coopté du conseil d'administration, au sein du jury du prix de thèse ou du comité de programme des Journées.

The board will be happy to appoint a member to serve in the scientific council of the foundation. The board will also be happy to have a representative of the foundation join its rank. From our experience, membership of the RePEc board is neither onerous nor particularly exciting.