ArchEc: an archive for RePEc, part 2

A funding application to the Banque de France Foundation for Economic Research

2021–01–08

This application comes on behalf of the RePEc community. It is for 3000 Euros to continue to work on a comprehensive archive of all the working papers that RePEc has links for, and to start working on the integration with the actual RePEc metadata. Thus, it is the continuation we referenced in the previous application of 2019–01–09. Then, as 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; full texts can easily disappear; and papers can be revised, but 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.

1.2 History

In 2019, the foundation paid €3000, roughly $3300 to the Open Library Society. The society is the organization representing RePEc. Since this was the first time the society got any money, it had to keep $500 in the bank account that the society opened to receive it to avoid recurring banking fees. Thomas set off to work with the remaining $2800.

An old adage goes: “Faster, better, cheaper — pick two”. Thomas picked the last two. He could not finish in 2019. Therefore, we did not apply for continuation funding in 2020. However, by this time, Thomas has done much of we set out in 2019. We will go into details in a separate document. Here we just want to illustrate the relative scale of the achievement with a simple example. According to https://www.nereus4economics.info/neeo\_intro\_press.html, in 2007 the European Union paid €988,104 for 50,000 metadata records, €19.76208 per record. As of 2021–01–06, Thomas reported 11205179 papers archived, with 1593774 PDF payloads, 1347598 of which are distinct. Thus, the foundation paid €0.00222618317 per PDF document. Now clearly a PDF document is much more than a metadata record, but even if we are to assume, as good economists, that they are the same, the Banque de France Foundation got an 8877 times bigger bang for the buck than the European Union. It was not a fast bang, but a big bang for sure.

1.3 Work

1.3.1 Unfunded work: ArchEc to NEP

This is work that is not conditional on funding success. Even if we do not get money this year, we will place a report on this effort in the next funding application.

Our archive is of no use if not used. However it must not be used heavily too early as all this unauthorized copying is problematic. A baby step is to use the stored PDF for NEP delivery. When NEP users ask for full-text, we deliver from ArchEc’s store. Only when no PDF full-text is in the store, then we try to deliver the text by full-text link. This implementation will also be important to test how many new working paper items we can actually reach as PDF. It may lead the NEP leadership to sanctions RePEc archives that have inaccessible PDF by excluding their contents from NEP.

1.3.2 Warranted work: RePEc data

As we alluded in the 2019 application, there was no organized effort to preserve actual RePEc data. There are dumps that Thomas created over the years. They are in an irregular shape. For the purpose of inspection, we have made them available at http://dumps.repec.org. From data captured on 2021–01–03, with 1225075285496 bytes, it is roughly half the size of the complete ArchEc archive, which stood at 2315752783081 bytes. The work under this application will split the contents into files to remove duplicates. It will look through the data for full-text items. These will be manually added to the ArchEc WARCs. It will collect the contents into tar.gz files per archive, rather than per dump date. We will publish some summary statistics. What we will not do is parse ReDIF data into RePEc records. We will not attempt any record-level work. How much time this will take is dependent on whether we can get a bigger server for additional disk space. Within the confines of the disk space we have it will be quite a cumbersome task.

1.3.3 Additional work: Internet Archive and Heritrix.

ArchEc uses wget, see https://en.wikipedia.org/wiki/Wget, to compose the WARCs. There are teething problems with the WARC function in wget, as set out in the final report. We know of only one alternative that natively writes WARCS. This is Heritrix, see https://en.wikipedia.org/wiki/Heritrix. It is produced by the Internet Archive. It is used for serious web archiving as performed, for example, by national libraries. Incidentally, Thomas knows a person running Heritrix at the French National Library. The Internet Archive recently started a program to archive open access contents from academic publishers. We are in touch with their staff. Predictably they warn us that Heritrix not meant to be run by amateurs, however good they maybe. But they have signalled they would send us their configuration they use for getting article full text from publishers’ sites. This will be useful, because getting such PDF full-text files is a very different job from doing broader crawls. We also hope that this configuration for Heritrix will solve the issue of accessing full text when only an intermediate page is available. Finally, working with Heritrix could prepare future work archiving sites from professional site in economics, such as scholarly societies.

Please note that we do not propose to work here on replacing wget with Heritrix. To do that, we will need to merge the existing WARCs with Heritrix WARCs. This may be (or may not be, we just do not know at this stage) a very expensive task.

In exchange from taking these tools from the Internet Archive, we will propose them to get all of our archived contents. Thus even if one day RePEc would not exist, the Internet Archive would have a copy of all our data. Once it is there, it is not likely to disappear in the foreseeable future.

1.4 Reporting

If there is no funding this year, we will report on the work in 1.3.1 and make a new application next year. If there is funding, we do not warrant that the work will all be done in 2021. We will make no further application until we have reasonable progress on both 1.3.2 and 1.3.3. In fact, any future funding application will heavily depend on what we have done there.

1.5 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. Clearly, the project is running on its own but on some occasion it needs external support. We do not pretend that the support from the foundation is an adequate compensation for the work. However, it means a lot to us. We are already proud that the Banque de France Foundation has become part of this trail-blazing piece of infrastructure. In economics, old research rarely looses significance. Future technology will allow people to ask questions directly into the intelligence collected in the millions of documents available. No future technology will be able to reproduce the documents once they have disappeared.

Part 2.

In this part, 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 rules on the governance web site. All the board does is pass resolution. The board has approved this application resolution dated 2021–01–08, available at http://governance.repec.org/resolutions.html

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 the work of 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 and ArchEc’s site at http://archec.repec.org have the foundation’s logo. We want to be seen as being supported by the foundation.

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. Membership of the RePEc board is neither onerous nor particularly exciting. We are not astonished that nobody has taken us up on our offer.