Montréal proposal

Thomas Krichel and Ivan Kurmanov

2002–10–09

Thomas Krichel
Palmer School of Library and Information Science
720 Northern Boulevard, Brookville
New York 11548–1300
USA
krichel@openlib.org
http://openlib.org/home/krichel

Ivan Kurmanov
MS B285
Chervyakova street, 8–151
Minsk 220002
Belarus
kurmanov@openlib.org
http://c2.com/cgi/wiki?IvanKurmanov

The latest version of this document may be found at http://openlib.org/home/krichel/work/montreal.html. It has benefitted from comments by Christopher F. Baum and Michael E.D. Koenig. Donald Arseneau provided advice on the use of his url.sty LaTeX package.

1 Introduction

0. This proposal comes from Thomas Krichel and Ivan Kurmanov, to the Soros foundation to support the creation of an Academic Contributor Information System (ACIS). The software will allow the construction of services where academic contributors identify themselves and create links to their contributions.

1. In conventional bibliographic data, the academic contributor is the author of a document that is described in a bibliographical dataset. However the concept of contribution is more general than the concept of authorship.

2. The identification consists in a set of personal data that a registrant supplies to the registration system. The link to the contribution consists of the handle of the document that has been contributed to and the contribution type. The contribution type is selected from a controlled vocabulary.

3. ACIS has been pioneered by the HoPEc system that is part of the RePEc project. Sections 2 introduces the RePEc collection and 3 deals with this legacy system. Readers who are familiar with them may skip these sections.

2 RePEc

4. 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 Open Archives Initiative (OAI) business model that distinguishes between data providers and service providers. Over 250 archives 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 up-to-date figures. RePEc 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 £129,000.

5. A conventional library is a collection of documents plus a user interface to search it. RePEc separates the task of building collection from the task of presenting the collection to users. The Open Archives Initiative (OAI) has more recently imitated this architecture.

6. There is another important difference between RePEc and conventional digital libraries. RePEc is not just a catalog of documents. In addition to document metadata RePEc collects data about institutions involved in the research process and researchers themselves.

7. 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 6,400 economics-research-related institutions in his dataset.

3 HoPEc

8. In 1999 the RePEc team, with support by JISC, 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 RePEc with their contact details, affiliation data and research data.

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

10. 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.

11. Research data is a list of references to the document items in RePEc. During the registration process, the system searches in the RePEc documents data for items which have a variant of the spelling of the registrant's name among the authors. The registrant 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".

12. 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.

13. 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.

14. A sample personal profile page is available at http://netec.mcc.ac.uk/adnetec-cgi-bin/gemini.cgi?submit=id&HANDLE=RePEc:per:1945-02-12:DAVID_FRIEDMAN. The 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 the

EconPapers RePEc service representing the same personal profile is http://econpapers.hhs.se/HoPEc/49575253485049506865867368957082736968776578.htm. Most importantly, we gather log data for all the person's paper, an example is 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 well 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.

15. Currently over 4,700 economists are registered. They have claimed authorship for over 34,000 publications. For more information on the existing service see http://authors.repec.org.

4 Motivation

16. Scholarly communication is a meeting place between authors and readers. It is best thought of as a market place where authors exhibit their works seeking wide dissemination and peer recognition. The principle form of output is the research paper.

17. It has been argued that the unrestricted access to the research papers enhances the dissemination of papers and that therefore researchers should be keen to take up this additional means to disseminate their works. Unfortunately this has not been taken up a great deal. It appears that the increased dissemination that self-archiving affords is only a weak incentive for scholars, in particular in the disciplines where there has been no pre-publication culture.

18. Building a scholarly communication system that is based on open access—as a result of the introduction of the Internet—is the task of authors, publishers and libraries. Libraries can operate institution-based archives. Publisher can offer free-access journals over the Internet. But to get the process really going, the really crucial role is the one of authors.

19. The benefit from open-access must be demonstrable to authors. Otherwise they will not take steps to make their works available. Thus, it must be demonstrated that the online access to the authors works has positive repercussion on the author himself. Such a demonstration can not be undertaken without a building a relational dataset between authors and publications.

20. The benefit from open-access must also be demonstrable to institutions, too. Thus, it must be demonstrated that the on-line access to the authors works has positive repercussion on the institution the author is affiliated with. Such a demonstration can not be undertaken without a building a relational dataset between authors' publications and authors and their institutions.

21. These relational datasets will contain authoritative handles for papers, authors, and institutions. It appears that creating one central authority with the task of providing these datasets is not realistic. Therefore it will be necessary to divide the total set of describables into several subsets. Of course these subsets may overlap. Each subset, a database, will be administered by a group who form the "authority" for the dataset.

22. It appears that the best sub-set division is the academic discipline. The reasons for this choice are obvious. Traditionally academics have been organizing themselves through disciplines. Most academics think of themselves as members of a discipline first and members of an institution second. However, a split of the total dataset on discipline lines is problematic because disciplines are fuzzy concepts. For example, is economic history a part of economics or a part of history?

23. A split of the whole of the academic lines on authority backgrounds will be in part an accident of history, as authorities are formed and take it upon themselves to collect data relating to a certain discipline. We expect that over time, authorities will be built to cover all disciplines, but we may be wrong, of course.

5 Partners

24. There will be three different authorities involved in the project. PhysNet will be represented by Eberhardt R. Hilf (University of Oldenburg). rclis will be represented by José Manuel Barrueco Cruz (University of Valencia). RePEc will be represented by Christian Zimmermann (University of Connecticut). None of the partners will be receiving funds under the present proposal. They will be indirect beneficiaries through the work that is conducted. They will make their data available through the Open Archives Initiative protocol for public metadata harvesting.

25. A steering committee will be formed for the oversight of the project. The representatives of the partners projects—as listed in the previous paragraph—will be ex-officio members of the steering committee and have agreed to participate. The Soros foundation will name two representatives to the steering committee. In addition, the following people will be approached to join the steering committee:

- Les Carr (University of Southampton)
- Steve Lawrence (NEC research laboratory)

- Michael Ley (University of Trier)
- Sergei I. Parinov (Siberian Branch of the Russian Academy of Sciences)
- Herbert Van de Sompel (Los Alamos National Laboratory)
- Simeon Warner (Cornell University)
- Jeff Young (OCLC)

26. Thomas Krichel will be acting as the overall project coordinator. He will award the funding and work with the project consultant on matters of protocol. He will not be paid by the project funds but may operate an expense account, within the limits imposed by the budget.

27. Ivan Kurmanov will be the principal consultant. He write the software, together with other developers that he may choose. He will report to Thomas Krichel on the delivery of the software. Ivan and his co-workers will be paid for their work by project funds.

28. All software will be released under the GNU Public license. But this is not all that it takes to produce software that is useful to others. The software needs to be well documented. Good documentation is required for other people to use it. In addition, the software needs to be carefully designed. Careful design implies that customizations and extensions are possible and don't take a rocket scientist. Good documentation and careful design take a lot of time and effort.

6 Stage One

29. In stage one, the project will create software to manage the creation and use of personal data that are in relation with document data, within an academic context. It will create an interface to capture and manage such data. It will not host the data, but install and maintain the software at sites hosted by the partners.

30. In stage 1, the software will work on an external, static sets of document data. The document data will be compiled by the partners. The input format for the document data will be the Academic Metadata Format (AMF) by default, but other XML formats may be supported by requests from the partners.

31. The interface to the personal data will appear as 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 document and affiliation information. As an optional feature, the software will allow people to give leave data about their research fields, as well as some personal information like a photo.

32. In order to provide for user-friendliness, the project will fully utilize the feedback received from the users of the HoPEc service. The software to minimize the number of required "clicks" and page reloads during the registration process and following updates.

33. At each partner site, the software will only appear in a slightly different way. Each page generated by the software will be produced in XML, and then filtered through an XSL stylesheet that will allow to customize the appearance of the page. The overall registration process, i.e. the fundamental procedure will be the same for all services. The process will be fully documented.

34. The project expects all the data generated by the software to be placed on the Internet with no restriction on usage, including commercial usage. It is understood that some inappropriate use may be made with the email addresses in the data. Unless there has been an incident of miss-usage, the data will be fully disclosed through version 2 of the OAI protocol for metadata harvesting (OAI-PMH).

35. On web sites, the email data will be hidden. Appropriate hiding will be investigated. It will probably involve using graphics rather than character data. A Perl module to implement the hiding of email addresses will be made available to user services.

36. Once a person is registered, she can opt for a semi-automated update of a profile. If this is enabled, the system will study every new document addition to the document dataset, to check if any of it's authors name matches that of any of the registered researchers. If the document should be added, it will inform the registered user, giving her the option to cancel the addition.

7 Stage Two

37. In stage two, the project will be extended from the authorship of documents to the authorship of citations contained in other documents. The system will scan citation data for the occurrence of the name of an author, and ask two questions. First, is this you who is being cited in this paper? Second, is this paper part of your research profile, i.e. the list of papers that is already available? We know that authors are very interested in obtaining data on citations to their works.

38. For PhysNet and RePEc, citation data is available through the Open Citation and CitEc projects, respectively.

For rclis, citation data could be gathered through collaboration with CiteSeer, but for the moment this is out of the scope of the proposal. It is an option that will have to be studied. The ACIS project will fund the conversion of metadata provided by the citation data sources to a common subset of the Academic Metadata Format, that will be used for input into the database.

39. ACIS will export the value-add citation data for usage by the contributing citation indices. A precise way of doing this will have to be agreed between participants.

8 Stage Three

40. In stage three, the interface with document archives will be set up. The inspection of metadata for non-downloadable papers will lead authors to want to make them available. Unfortunately archiving papers is not the task of an ACIS service. It would be quite foolish to add an uploading facility. Thus, the objective must be to improve working with existing archives, rather than replacing them.

41. The applicants will be working on a general protocol for data updating between trusted parties on the Internet. When an author has uploaded a paper to a participating archive, she can choose to add the paper to her research profile, as well as to the research profile of any other co-authors who have opted for semi-automatic updates.

42. The update will be performed in real time. The document archive will be aware of where the machine that houses the author data is. This could be done with a DNS query, for example. It then sends a request to find if a certain author handle, as supplied by the submitter of the paper is valid. If it not valid, the submitter will be warned and the submission of the handle is invalid. If all quoted identifiers are valid, the new document record is exported—using the SOAP protocol—to the ACIS service, and an update of the author profiles is performed accordingly.

43. As a proof of concept, the protocol will be implemented at the Economics Working Paper Archive at Washington University of St. Louis as a prototype system.

44. A module to handle the same functionality will be provided to the eprint software. Support for integrating it into arXiv will also be provided. However the main outcome here will be a general protocol that is applicable in this circumstance and that may be used widely in situations where different participants update a relational database is a cooperative way.

9 Stage Four

45. In this last stage, the project will work on calculating evaluations of the impact of registered persons. The personal data that the partners have accumulated will be joined to impact measures of the documents associated with the documents.

46. The LogEc project, which is part of RePEc, has already done pioneering work is this effort for the RePEc data. But the measures that it proposes are very simple.

47. To get this to work, impact measures of documents must be defined. There are, of course, many ways in which the impact of a document can be defined. We can count instances such as downloads, abstract views, inclusions in certain collections, citations by other documents and

48. The project will work on a descriptive model of services and service incidents and an evaluative model describes which basic evaluative methods are usable. Within an evaluative method, data from system-wide incidents is translated by a function into a number, which is basically an expression of how well the contributor does with respect to the chosen criterion.

49. There is no hope to find a descriptive syntax that encodes all evaluative methods that one may potentially be interested in. The project will aim to identify the best evaluative methods, and find ways to encode them. A good evaluative method

- can actually be meaningfully explained to users;
- can be calculated from the data that the partners have;
- can be displayed in a visually attractive way;
- is not subject to moral hazard or adverse selection.

50. The project will help the partners to build web services that calculate and display evaluative data. It is premature to try to set this out at the time of writing.

10 Budget

51. The expected time line is as follows. Stage one is expected to be completed in five months, stage 2 will take four months, stage 3 will take seven months and stage four will take six months. This gives a 22 month total time. We will add two spare months for security. Thus the project will start on 2003–01–01 and end 2004–12–31.

52. Ivan Kurmanov will be paid a such of $1,000 per month, which he will hire labor from to form a small team to support him. In addition, he will be paid an additional $3,000 on completion of every stage. However, if progress on work is not sufficient, on each of the dates when a stage is supposed to end, Thomas Krichel may—subject to the agreement of the steering committee—hand over the work to another consultant.

53. Thomas Krichel will be awarded a $5,000 expense budget line. This may be used to cover travel expenses, and attendance at conferences and other meetings to promote the project. They may not be used for personal items or computer hardware.

54. In addition, there will be a $7,000 account for exceptional work to be done by the partners who build the datasets. Any money that will be disbursed under this post is subject to agreement of the steering committee.

55. Thomas Krichel will name the organization that administers the funds. $2,000 will be set aside to compensate the organization for the costs associated with the financial administration. With the agreement of the steering committee, Thomas Krichel may entrust another organization with the financial administration. Such a change would result in a prorata compensation over the time the financial administration has been carried out for.

56. Here is the budget in tabular form, in United States dollars.

item		*cost*
software consultants, basic monthly	1,000	
times 24 months		24,000
software consultants, bonuses per stage	3,000	
times 4 stages		12,000
project director expenses		5,000
special projects for datasets		7,000
administrative expenditure		2,000
total		50,000