PROMOTING OPEN ACCESS TO SCHOLARLY DATA: A CASE STUDY OF THE ELECTRONIC THESIS AND DISSERTATION (ETD) PROJECT AT THE SIMON FRASER UNIVERSITY LIBRARY

Ian Song

W.A.C. Bennett Library, Simon Fraser University, 8888 University Drive, Burnaby, British Columbia, Canada.
Email: isong@sfu.ca

ABSTRACT

Scholarly data, such as academic articles, research reports and theses/dissertations, traditionally have limited dissemination in that they generally require journal subscription or affiliation with particular libraries. The notion of open access, made possible by rapidly advancing digital technologies, aims to break the limitations that hinder academic developments and information exchange. This paper presents the *Electronic Thesis & Dissertation* (ETD) Project at the Simon Fraser University Library, British Columbia, Canada， and discusses various technological considerations associated with the Project including selection of software, capture of metadata, and long-term preservation of the digitized data. The paper concludes that a well-established project plan that takes into account not only technological issues but also issues relating to project policies, procedures, and copyright permissions that occur in the process of providing open access plays a vital role for the overall success of such projects.

Keywords: Open Access; Scholarly Data; Electronic Thesis and Dissertation; Digitization; Copyright; Simon Fraser University.

1 INTRODUCTION

Open Access (OA), as a new approach to scholarly communication, has attracted people's attention since Budapest Open Access Initiative (BOAI) was formally released by Open Society Institute (OSI) in 2002. The purpose of the BOAI meeting in 2001 was to make an effort internationally to make research articles in all academic fields freely available on the Internet. BOAI has been regarded as a monument of the Open Access movement, and it has been one of the hottest topics among scholarly communities, libraries, publisher, and the general public since then.

"Open Access was named one of the top science news stories of 2003 by Nature, Science, The Scientist, and The Wall Street Journal" (Koudinov & Suber, 2004). OSI Chairman George Soros donated 3 million dollars to support OA projects around the world from 2002 to 2005 (BOAI, 2005). Many of the original signatories of the BOAI are the leaders and advocates of the OA movement (Wikipedia, 2006). Over four thousand individuals and over three hundred organizations have signed the initiative so far (BOAI, 2006).

OA promises to make scholarly data or literature more accessible and more affordable and brings a hopeful solution to the scholarly communication crisis that has significantly impacted scholarly communities, libraries, and library readers for decades.

1.1 What Is OA?

There are many definitions of OA, causing much confusion. BOAI's definition is considered one of the most authorative (along with Berlin and Bethesda) (Suber, 2006). The following elements should be clearly defined:

1 Literature Type
OA literature should include scholarly peer-reviewed articles, non-peer-reviewed preprints, and any scholarly writings for which authors do not expect payment but not include non-scholarly writings, such as novels, newspapers, and poems.

2 Media
OA literature should be in a digital format and on the Internet where people can freely access it without legal or technical barriers.

3 Users' rights
For OA literature, a user should be able to read it, download it, copy it, distribute it, print it, and handle it for any other lawful purposes.

4 Copyright
The copyright holder can be the author of an OA item or the publisher of an OA article. The copyright holder keeps the copyright of an OA item and decides whether an academic article or writing is open access or restricted.

There are many debates about OA. Peter Suber recommended BBB 's (Budapest-Bethesda-Berlin) definition (Suber, 2004). ARL (Association of Research Libraries) considers OA as an alternative to the traditional subscription publishing model, and some people regard OA Self-Archiving as a supplement to the toll-based primary literature at this time, not substitutes for it (ARL, 2004 and Eprint, ?).

1.2 Rationale of OA

Scholars, libraries, readers in the scholarly communication system, along with many smaller publishers, have been suffering from the so called scholarly communication crisis for a few decades, which made libraries cancel subscriptions on peer-reviewed journals due to an increase of over 270% of the price of peer-reviewed journals from 1986 to 2003 (Kyrillidou, 2004).

A small and decreasing number of highly profitable publishers are the only parties in the system that win the game as more and more mergers of publishers put them in a good position and give them more power to control important factors, such as price, format, and distribution, in scholarly publishing. Many smaller publishers and publishers outside of STM (science, technology, and medicine) have been hurt by the mergers and rising prices of the large players. There have been cuts to journals and monographs in the social sciences and humanities, for example, to pay for the rising prices in the sciences. Scholarly information that originally was a public good, funded by taxpayers' dollars, was changed into a relatively inelastic, special commodity (Edwards, 2003).

Some people say it is time for scholars, librarians, and readers to fight back, but in the author's opinion, it is time to find solutions to the scholarly communication crisis. Some university libraries have been trying to develop different solutions to the crisis, such as creating buying consortia or collaborative bodies by sharing licenses. However, this kind of effort seems not to resolve the problem completely. Reform or restructure of the foundation of the current scholarly communication model becomes more critical than ever.

Academic research articles and other academic writings from public-funded research have been considered as a public good, and they should be available for wider public access (Edwards, 2003). The following quote from John Willinsky's new book "The Access Principle" clearly states the access principle:
"A commitment to the value and quality of research carries with it a responsibility to extend the circulation of such work as far as possible and ideally to all who are interested in it and all who might profit by it." (Willinsky, 2006)

OA has been full of potentials and promising ideas since it was formally introduced at Budapest Open Access Initiative as a new paradigm for dissemination of scholarly research, and OA has brought a hopeful solution to the serials crisis. OA will remove two major barriers that prevent users from accessing academic materials: the price barrier and the permission barrier (Suber, 2004).

1.3 How to achieve OA

OA can be achieved in two ways: OA journals (Gold Road) and Self-archiving (Green Road). Self-archiving is also called “OA Repository” or “OA Archives”. Generally speaking, it can be either “Institutional repository” or “Subject (disciplinary) repository”. This paper will be focusing on “Institutional Repository”.

This paper will focus on OA archives and use the Electronic Thesis and Dissertation Project at Simon Fraser University Library as a case study to further discuss the new paradigm as a solution to the scholarly communication crisis.

2 INSTITUTIONAL REPOSITORY (IR)

IR as a permanent, institution-wide repository of diverse locally produced digital works has been widely used in academic institutions since 2002. ARL conducted a survey on IRs in January 2006, and its results showed the current status of IRs in ARL. 43% institutions in ARL have operational IRs and 35% are planning to implement IRs in 2007 (ARL, 2006).

Again, there are many definitions for institutional repository. One of the best, in my opinion, is the definition from Raym Crow**,** a SPARC (Scholarly Publishing & Academic Resources Coalition) senior consultant: “digital collections capturing and preserving the intellectual output of a single or multi-university community” (Crow, 2002).

In an Institutional repository, you may find preprint articles, peer-reviewed articles (postprints), theses, research data, presentations, and other research related materials. However, IR is not an organizational records management system.

One of the most important features of IR is its interoperability, which allows different systems to work together to accomplish a common task and share data (Harnad, 2001). OAI (Open Access Initiatives) is a tool to harvest the metadata from each repository (data provider) into one virtual archive via its protocol “Open Access Initiatives Metadata Harvesting Protocol” (OAI-MHP).

One good example is the Canadian Association of Research Libraries’ (CARL) Institutional Repositories Pilot Project Harvester. This Harvester is the search service for the CARL Institutional Repositories Pilot Project; it aggregates materials from each of the participating Canadian institutions, allowing users to seamlessly search all of the repositories at once, using one common point of access (http://carl-abrc-oai.lib.sfu.ca/service_provider.php).

The *Directory of Open Access Repositories (OpenDOAR)* and *Registry of Open Access Repositories* collect IR applications around the world, and they are good tools for investigating the development of IRs.

The results of research conducted by Lynch (Lynch, 2005), CARL Institutional Repository Survey (Shearer, 2004), and ARL’s SPEC Kit (http://www.arl.org/pubscat/pr/2006/spec292.html), show that the collection of theses/dissertations is one of the major contents in the IRs in Canada and the US.

3 ELECTRONIC THESES AND DISSERTATIONS (ETD)

Traditionally, theses/dissertations as an important part of scholarly data are physically housed in the libraries that are affiliated with the universities where the authors pursued their advanced degrees. Theses and dissertations contain valuable research information that summarized the entire research interests and achievements of their authors. Starting in the late 1980s, the great success of digital technology and personal computers made electronic theses and dissertations possible. Since the mid-1990s, the electronic theses/dissertations movement gained great momentum, and ETD became more and more popular in the United States, Canada, and some European countries （Fineman, 2003）. ETD makes theses more accessible than ever.

Most ETDs have some restrictions and limitations that prevent users from accessing them. For example, the theses and dissertations at ProQuest/UMI (University Microfilms International) are fully accessible to their home university users without charges. Other users can get only 24 pages in a thesis for free and have to pay if they want to read a whole thesis. In Canada, Theses Canada, affiliated with Library and Achieves Canada, working with UMI, "Indexes over 250,000 Canadian theses and dissertations held by the National Library of Canada. Also, free access is provided to 46,000 full text Canadian theses and dissertations covering 1998-2002." (http://library.humboldt.edu/~rls/theses.html) SFU is one of the 58 participating universities. This is a great effort to promote open access to academic theses/dissertations, but unfortunately, it covers a short period of time, and most of theses in the portal are not full text.

NDLTD model (Networked Digital Library of Theses and Dissertations) is a mixed model in terms of access. Some of its holdings are fully open; some are restricted; some are fee-based. DNLTD is an OAI-based union catalogue. The metadata records on all the sites (institutions) should be described with similar or common metadata standards so that the records from different institutions can be retrieved more efficiently and effectively. Since 2002, a similar model has gradually become the main stream of ETD management. Many universities are using institutional repositories as content management systems (more information about IR software can be found at BOAI's website), most of which are open source systems to manage ETDs. Each institution as a data provider has its own collection of theses and metadata on each site, which can be harvested by a service provider.

In addition to the models above, some theses authors put their own theses on their own website or their departments' website and provide free access.

In the following sections, the Electronic Thesis and Dissertation Project in the Simon Fraser University IR will be discussed as a case study.

4 ETD AT SFU

4.1 SFU

Simon Fraser University (SFU) is a young, medium sized comprehensive university and has offered undergraduate, masters, and PhD programs since 1965. It has over 25,000 students, including over 3,000 postgraduates, on three campuses. From 1965 to 2004, over 80 thousand degrees were conferred, including over 17 hundred PhDs and 10 thousand Masters. More than 600 theses are submitted each year.

4.2 Planning

In 2003, SFU Library planned to create electronic theses/dissertations based on past experience in digitization of the materials in its special collection. The theses between 1998 and 2002 were already in Theses Canada, the rest were the retrospective theses (1966-1997) and new theses after 2002. "UMI approached us with its proposal of converting SFU theses to digital format and providing MARC records, but UMI would retain a copy in its database and provide free access to SFU users only." (Mundle, 2006) After investigating possible solutions and estimating the cost, we decided to do the conversion ourselves. One of the major advantages was that wider access to the theses could be provided. Non-SFU users also can access SFU scholarship without any charges.

In the plan, the SFU ETD project will be completed within two years, and over five thousand retrospective and new theses will be digitized. In the meantime, former graduates will be contacted in order to obtain their permission to put their theses in IR.

By the end of this year (2006), the digitization of over 5 thousand theses will have been completed.

4.3 SFU's IR

DSpace from MIT and HP was chosen to manage the digital theses collection, which captures, stores, indexes, preserves, and redistributes an organization's research data. DSpace is an open source and interoperable software. It also complies with the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), which can harvest metadata from each repository based on the open standard HTTP and XML, bring all the metadata from different repositories together, support Dublin Core schema, and provide the base for coming Dspace federation in future.

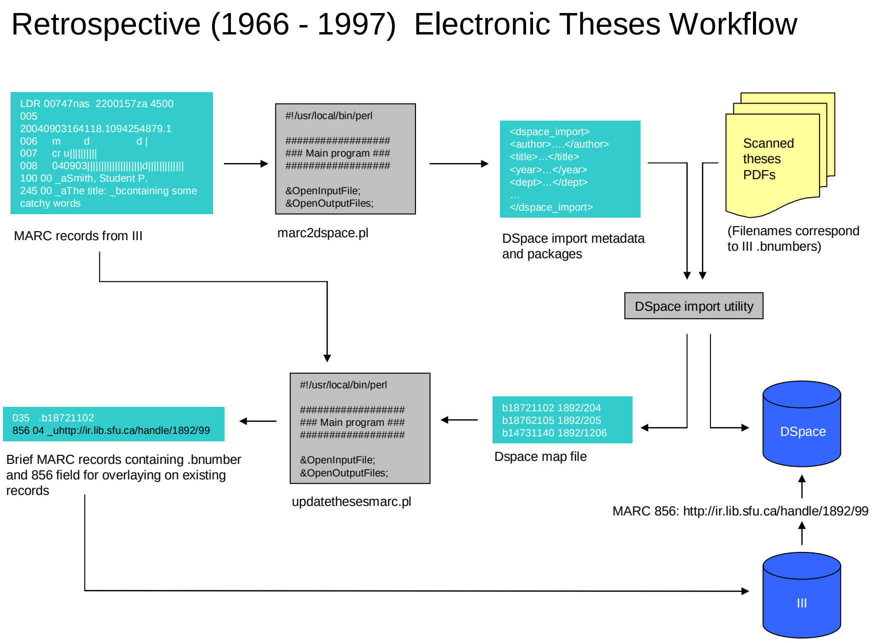
SFU started its IR in 2004. It now has 9 communities and over 1,500 documents, including postprints, research papers, theses, conference presentations, and other documents.

SFU's IR has its guidelines and polices regarding the contents, privacy, and its collection (http://ir.lib.sfu.ca/policies/index.jsp).

4.4 Digitization and metadata

In 2004, the ETD project at SFU started with large scale digitizing of retrospective printed theses and theses on microfiche. At the same time, new theses have also been digitized before they are sent off to Library and Achieves of Canada. Since then, over five thousand theses have been digitized and formatted to searchable PDF files.

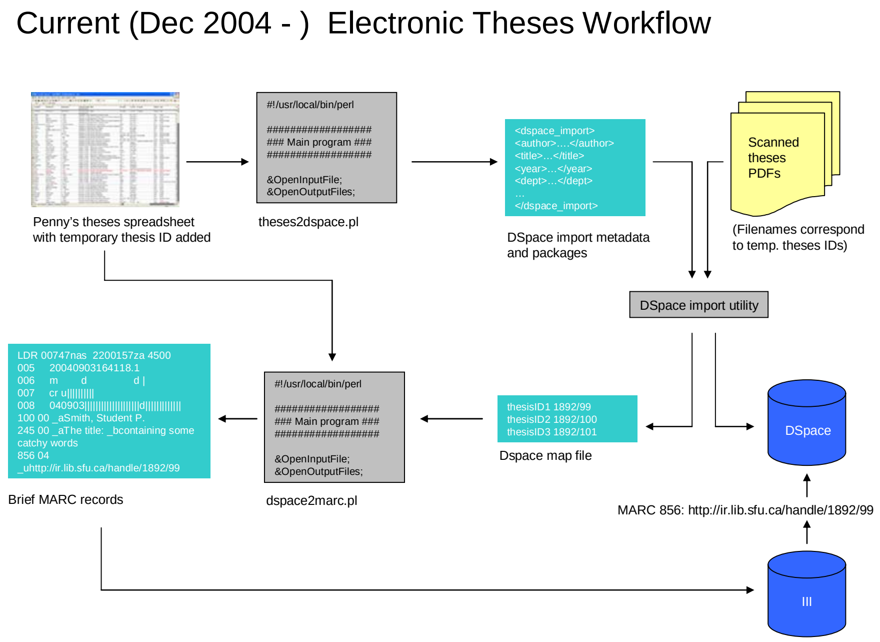
All the retrospective theses have their MARC records in the library's catalogue, which can be converted into the DSpace's Dublin Core metadata format via a Perl script as shown in Figure 1:



Retrospective (1966 - 1997) Electronic Theses Workflow

Figure 1 * (Courtesy of Mark Jordan)

Since new theses do not have their MARC records, a spreadsheet, which contains required metadata information in Dspace, is created either by the Theses Office or students themselves via online submission. Another script converts the metadata information on the spreadsheet to DSpace metadata format. In this way, a thesis PDF file and its descriptive medadata are imported into DSpace. This script also creates a brief MARC record for each thesis in the catalogue system as shown in Figure 2.


Current (Dec 2004 -) Electronic Theses Workflow
Penny's theses spreadsheet with temporary thesis ID added
theses2dspace.pl
DSpace import metadata and packages
Scanned theses PDFs
(Filenames correspond to temp. theses IDs)
DSpace import utility
Brief MARC records
dspace2marc.pl
Dspace map file
DSpace
MARC 856: http://ir.lib.sfu.ca/handle/1892/99
III

```
#!/usr/local/bin/perl

##################
### Main program ###
##################

&OpenInputFile;
&OpenOutputFiles;
```

```
<dspace_import>
<author>….</author>
<title>…</title>
<year>…</year>
<dept>…</dept>
…
</dspace_import>
```

```
LDR 00747nas  2200157za 4500
005    20040903164118.1
006    m      d          d |
007    cr u|||||||||
008    040903|||||||||||||||||d||||||||||||
100 00 _aSmith, Student P.
245 00 _aThe title: _bcontaining some
catchy words
856 04
_uhttp://ir.lib.sfu.ca/handle/1892/99
```

```
#!/usr/local/bin/perl

##################
### Main program ###
##################

&OpenInputFile;
&OpenOutputFiles;
```

```
thesisID1 1892/99
thesisID2 1892/100
thesisID3 1892/101
```

Figure 2 * (Courtesy of Mark Jordan)

*Note: b number is a unique ID for each item in III that is an integrated library system developed by Innovative Interface, Inc.

4.5 Right management

4.51 Copyright, Partial Copyright License, and permission issues

Copyright is a double-edged sword. On the one hand, it protects the original works of authorship and gives the copyright owner exclusive rights to reproduce the works. On the other hand, it limits how libraries and readers may use the materials the library or its organization already paid for.

Traditionally, the author of a thesis gives the university library permission in the Partial Copyright Licence to lend the paper thesis to library users, make partial or single copies of the thesis for its users, and lend it to other universities in response to interlibrary loan requests.

After consulting with the university legal counsel, based on Canadian *Copyright Act*, section 29.1 concerning the "fair dealing" exception, the Partial Copyright Licence was revised in order to facilitate the online thesis service in 2004, in that theses authors further granted permission to the university to make digital copies for use in university's circulating collection.

The most challenging issue in the project is to obtain permission from former graduates from the past 40 years for the library to digitize their theses and put these digitized theses online. Over five thousand authors have been, or will be, contacted before these electronic theses are archived in the IR.

4.52 Privacy

The personal, private, and confidential information in a thesis is well protected. Personal signatures on the

Partial Copyright Licence page, Approval page, and some correspondent letters are removed. Other personal information, such as email address, is also removed from theses, in order to protect the author's privacy.

4.6 Access

SFU's ETD collection in the IR has been open to the general public since 2004. In its IR, users can browse by authors, titles, and publication date. At the same time, keyword search in title, author, abstract, and other fields that are described in the metadata is made available.

To provide easy access to SFU online theses, 865 fields have been added to the theses' MARC records so that users can follow the URL links in the catalogue search result page to their records in DSpace.
Theses and dissertations at SFU institutional repository can be obtained in the following three ways:

- Searching the SFU catalogue and other union catalogues, such as British Columbia's OutLook OnLine,
- Searching SFU DSpace database,
- General Internet Search Engines, such as Google, and others.

The PDF files in the IR are full-text keyword searchable, once they are opened.

4.7 Benefits

The major benefits from this project are:

- Self control of digitization and related polices,
- Wider access to the SFU theses collection,
- Accumulation of experience on promoting OA, including rights, IR, and metadata management.

5 CONCLUSION AND CHALLENGES

Since BOAI, Open Access as a new publishing paradigm has attracted people's attention. Over ten speakers at the 20th CODATA International Conference spoke about OA and OA related topics, and many participants showed their great enthusiasm.

The OA movement will become a great monument and change the traditional publishing model significantly. However, it is not realistic to abandon the traditional publishing and subscription model right away. Many people believe that OA will coexist with the traditional publishing model, and both sides will adopt each other. Libraries and readers will benefit from either pure OA or a hybrid of both.

Rich experience has been obtained from our ETD project. There are still some challenges that need to be addressed, such as permission issues (already discussed above) and long-term preservation of ETDs.

The goals of digitization have been controversial for a long time. Promoting access is certainly one of major goals of digitization, but preservation is really questionable except for the materials that are in measurable danger.

LC, OCLC, ARL, CARL, and many other organizations have developed some frameworks and strategies for digital preservation. However, it is still too early to regard digitization as an appropriate long-term preservation strategy (Universal Preservation Format, 2006).

In the SFU ETD project, the original copies of paper theses or microfiche will be kept after they are scanned. We do not have any weeding plans for these theses. The PDF (Portable Document Format) file for each thesis is kept, but we do not keep TIF files in light of the potential cost of archiving. As born-digital theses will become more popular and acceptable in the future, their preservation has to be seriously considered in advance. Can we fully reply on PDF, one of the most popular file formats in preserving digital products? No one can tell whether a PDF file will still be readable and Adobe will continue to support PDF in the next two decades. The ideal format for digital products should be non-proprietary and independent of environment (hardware and software) or interoperable.

Preservation is a process, not a one-time deal. The process may involve regular data backup, data migration and changes of preservation media.

.

6 REFERENCES

Association of Research Libraries (2004) Framing the issue: open access. Retrieved September 8, 2006 from the World Wide Web: http://www.arl.org/scomm/open_access/framing.html

Association of Research Libraries (2006) Publication of SPEC Kit 292: institutional repositories. Retrieved September 21, 2006 from the World Wide Web:http://www.arl.org/pubscat/pr/2006/spec292.html

Budapest Open Access Initiative (2005) Open access projects supported by the OSI information program as of April 2005. Retrieved September 8, 2006 from the World Wide Web: http://www.soros.org/openaccess/grants-awarded.shtml

Budapest Open Access Initiative (2005) View signatures. Retrieved September 8, 2006 from the World Wide Web: http://www.soros.org/openaccess/view.cfm.

Crow, R. (2002) The case for institutional repositories: a SPARC position paper. Retrieved Oct 10, 2006 from the World Wide Web: http://www.arl.org/sparc/IR/ir.html.

Edwards, R. & Schulenberger, D. (2003) The High cost of scholarly journals (and what to do about it). *Change 35* (6): 11-13

Eprint (?) Self-Archiving. Retrieved September 8, 2006 from the World Wide Web: http://www.eprints.org/openaccess/self-faq/

Fineman, Y. (2003) Electronic theses and dissertations. *Library and the Academy 3* (2), 219-227.

Harnad , S. (2001) The Self-Archiving initiative. *Nature 410* (2001), 1024-1025.

Koudinov, A. R. & Suber, P. (2004) Open access, a breakthrough for science that every neuroscientist should know about. Retrieved September 11, 2006 from the World Wide Web: http://neurobiologyoflipids.org/openaccess/sfn2004.html

Kyrillidou, M. (2004) Serials trends reflected in the ARL statistics 2002-03. *ARL Bimonthly Report 234* (June 2004). Retrieved September 11, 2006 from the World Wide Web: http://www.arl.org/newsltr/234/serials.html

Lynch, C.A. & Lippincott, , J. K. (2005) Institutional repository deployment in the United States as of early 2005. *D-Lib Magazine11* (9). Retrieved September 11, 2006 from the World Wide Web: http://www.dlib.org/dlib/september05/lynch/09lynch.html.

Mundle, T. (2006) Digital retrospective conversion of theses and dissertations: an in house project. Retrieved September 1, 2006 from the World Wide Web: http://adt.caul.edu.au/etd2005/program.html.

Shearer, K. (2004) CARL institutional repository project: survey result-summer 2004. Retrieved September 19, 2006 from the World Wide Web:
http://www.carl-abrc.ca/projects/institutional_repositories/pdf/survey_results_2004-e.pdf

Suber, P. Budapest Open Access Initiative: frequently asked questions. Retrieved September 11, 2006 from the World Wide Web: http://www.earlham.edu/~peters/fos/boaifaq.htm#openaccess

Suber, P. (2004) Welcome to the SPARC open access newsletter, issue #77. Retrieved September 11, 2006 from the World Wide Web:http://www.earlham.edu/~peters/fos/newsletter/09-02-04.htm

Universal Preservation Format (2006) Ultimate goals concerning digital preservation. Retrieved September 3, 2006 from the World Wide Web: http://info.wgbh.org/upf/survey/survey06.html.

Wikipedia (2006) Budapest Open Access Initiative. Retrieved September 22, 2006 from the World Wide Web: http://en.wikipedia.org/wiki/Budapest_Open_Access_Initiative.

Willinsky, J. (2006) *The Access principle* Cambridge, MA: The MIT.