Analysis of Faculty Citation Behavior in the Electronic Age: A Study of one Institution's Recent Publications

Jim O'Donnell
Geological & Planetary Sciences Librarian
Caltech Library System
California Institute of Technology
Pasadena, California 91125
(jimodo@caltech.edu)

Everything's available electronically now! How true is that adage? In October of 2000, the Caltech Library System set out to evaluate this conceit, and gathered data to see just how close to true that statement is. Our goal was to discover just how much of the literature cited by Caltech's faculty was, in fact, available electronically.

INTRODUCTION

In the Autumn of 2000, the Caltech Library System was faced with a situation that revolved around one Theoretical Physics Professor's belief that everything that anybody needed was available on the web, and that the main library on campus was just a warehouse for books and serials that nobody was using anymore. Since this specific physicist happened also to be a highly placed official of the Institution, and since he didn't seem to care much about faithfully collected use statistics, we decided to face this challenge head-on by actually asking the question: Well, how much *is* available on the web?

For a physicist, 'Everything's available on the Web' is actually closer to true than for almost any other field of science: the Physics Preprint ('xxx') server (formerly at Los Alamos National Laboratory and now at Cornell University) was designed for physicists to share their work in preprint form immediately: the fact that it doesn't serve that purpose for all other sciences was a point that we hoped to make. Add to this the fact that the Institute of Physics (IOP) and American Physical Society (APS) -- both prolific journal publishers – have been leaders in the scanning and mounting of their journals' entire runs online, and you can understand that a physicist might assume that the same was true throughout the scientific world.

The Questions

In order to test this theory, we identified the following questions to answer:

1. What is *really* available electronically?
2. What proportion of the material that faculty cite is, in fact, electronically available?

3. Not constrained by Caltech's collections, what materials were available anywhere electronically on October 1, 2000?

Implicit in the first question is "What volumes of a journal are available online?" *Tectonophysics* is available online, but only beginning in 1999. The previous 308 volumes are not. This is the norm, rather than the exception. While the IOP, the APS and the American Chemical Society have scanned and loaded their journals back to volume 1, and J-Stor has full back runs of its journals, the fact is that commercial publishers are not scanning, indexing and loading their back runs: they're making available what they have in electronic format as a by-product of paper publishing. Our counting therefore made that distinction. Also, we cast a wide net: we included papers that were posted on an author's website.

We had to set an cutoff arbitrary date for ourselves, so October 1, 2000 was established. This was necessary because, as you'll see from the numbers included here, Caltech authors are decidedly prolific publishers.

The Plan

1. Identify the three most recently published refereed publications by Caltech tenure-track faculty.
2. Analyze the references in these articles to determine their sources.
3. Determine the electronic availability of those sources.

We worked with a single caveat: there was to be no duplication among citations. If a publication had already been added to the list because another author was also Caltech faculty, the Librarian found an earlier publication for that author. Multiple authorship might have really complicated matters, but we decided that we would only analyze each publication once.

Caltech is a small science and technology university with a very large reputation. At the time of this study, there were 285 tenure-track faculty in six divisions (Biology; Chemistry and Chemical Engineering; Engineering and Applied Science; Geology and Planetary Science; Physics, Mathematics and Astronomy; and Humanities and Social Sciences). There are about 2000 students, of whom 55% are graduate students.

Data Collection

The work was done by literally everybody on the Library System staff. Design of the project was done at the Assistant University Librarian level, and research of publications was done by the Librarian subject bibliographers. Implementation and design of the various tools used were done by the Head of Circulation and Document Delivery and a member of the Library System's Information Technology Group, and input and analysis were done by staff members of the Circulation and Document Delivery Group and the Technical Processing Services Group. Paging, photocopying, interlibrary loan -- all were among the tasks that were necessary to make this project happen.

The bibliographic data was collected into Microsoft Access by subject bibliographers. A Web of Science (WoS) search then produced a list of citations, which were emailed to a Circulation staffer who inserted them into a Microsoft Excel spreadsheet. The Excel citations were linked to the Access database by a unique record number for each paper.

To discover the electronic availability of the cited references, the bibliographers first consulted our Library Catalog, then surfed the internet looking for journals, authors' posted papers, and print-to-web conversions of books and other materials.

Sources of Data

Our primary source of data was Web of Science, which we have loaded on campus back to 1979. It is known for its speedy indexing, cited references, and a standardized format that we adopted as our own. It is a good source for most Caltech research fields, especially those whose publications are primarily in the journal literature. Geology (at Caltech, at least) is definitely among these. WoS is not good for computer science and some engineering fields that depend on conference proceedings for a lot of their publishing. It is decidedly poor for humanities (although we do have the Arts & Humanities Citation Index loaded) because Caltech's humanities professors tend to publish in books with distributed reference lists. Any publication that wasn't in WoS (and there were a substantial number) had to be entered by hand, or sometimes by a laborious scanning/optical character recognition/parsing of citations process that most of us didn't have to undertake.

Other sources of data included faculty websites (which we discovered had two interesting facets: (1) the number of papers mounted on them, and (2) the number that are not kept up to date). Besides these, various bibliographers used MathSciNet, INSPEC, and other bibliographies. For the record, of the 38 faculty that this author researched, only 3 were not in Web of Science – all earthquake engineers. For these I used Earthquake Engineering Abstracts Online and faculty websites to collect information.

Microsoft Access

Access is a relational database program that allows one to have many different tables linked to one another, and data can thus be analyzed using queries that are constructed by individual users. Our database has six tables. The fields are delineated in Table 1.

Table 1. Access Tables. Table name in bold.

Faculty

- Name
- Division
- Option (Caltech doesn't have 'departments')
- Faculty ID (unique record #)
- Librarian's (the data collectors) initials.

Publications

- Publication ID
- Material type
- Full bibliographic citation
- Parsed citation (author, article title, volume, source title, year, pages)

Cited Sources

- Title as cited in ISI
- Full Title
- ISSN
- Online (yes/no)
- Year
- Publisher
- URL
- ISI Title Abbreviation
- Comments

Cited References

- Cited References ID
- Material Type
- Publication ID (links to Publications Table)
- Volume
- Title
- Year
- Electronic Access (yes/no)
- Comment

Material Types

- Book
- Book Chapter
- Conference Paper
- Government Document
- Interview
- Journal Article
- Newspaper
- Other
- Patent
- Review Article
- Standard
- Technical Report
- Thesis
- Working Paper

Material type was important because it allowed us to demonstrate what kinds of sources our researchers are using. It's not enough to say "10,000 periodical citations" when you need to justify newspaper subscriptions and government document depositories, which are frequently forgotten by folks who focus on scholarly journals.

Problems encountered – desktop technology

Access was originally chosen because it allows multiple simultaneous use of the same database. However, we found that it was easier to manipulate files in Excel, with which most of the Librarians had at least some experience, than in Access, which was really only well-known by a few staffers.

We discovered very soon that the Access files were enormous, and that our staff workstations couldn't handle the load of even the main Publications Table, 11 megabytes of data. So a lot of the data were collected into Excel, and then transferred into Access for manipulation, a relatively simple process since they're both part of the Microsoft Office Suite.

We didn't have old computers: we had just never needed this much computing power before. This meant that a good portion of the work had to be done by one or two people who were, essentially, a bottleneck (however unavoidable) in the project's flow. Still, we managed to finish the majority of the work in about two months. One positive result of this: everybody on the Library System staff now has more powerful computers.

Campus findings

1. There were a total of 842 faculty publications, stretching as far back as 1982 to find three for each faculty member. 62% were published in 1999 and 2000.
2. A significant number of citations in these papers were to items published as far back as 1970. See Appendix 1 for details.
3. The citation statistics indicate an overwhelming dependence on the journal literature. Of 36,064 citations, 23,855 (roughly 66%) were journal articles. Another 5586 (about 15%) were books. The next largest number, Other, with 2348 items, is an amalgamation of all items that don't fall into any of the other categories. See Table 2.
4. 54% of the journals cited had an online presence (meaning that some part of the journal run was in electronic form). However, when analysis was complete, we found that only 38% of all citations were available online.

Table 2. Campus: citation statistics

Journal article	23855
Book	5586
Book chapter	1678
Conference Paper	1400
Technical Report	473
Thesis	370
Newspaper	162
Government Doc	107
Monographic series	68
Patent	9
Standard	8
Other	2348
Total	36064

Geology & Planetary Sciences Division Results

The author then worked with a subset of the campus database to glean information about his Library's primary population, the Division of Geology & Planetary Sciences (GPS). The next tables show these findings.

The Division, the smallest of the five science and engineering divisions, has 35 faculty. Of their 105 publications, 104 were journal articles (2 of which had no citations), and the remaining one was a book chapter. In these publications, there are a total of 4,670 citations.

Table 4. GPS: Publication year of Faculty papers

2000	72
1999	19
1998	7
1997	5
1996	2
Total	105

Compare this to 1982-2000 for the campus as a whole.

There were 4670 citations in those 105 papers, which broke down by material type as shown in Table 5.

Table 5. GPS: Citations by type

Journal Article	3601
Conference Paper	439
Book Chapter	242
Book	176
Thesis	87
Technical Report	68
Monographic Series	39
Other	18
Total	4670

Of these 4670, we found that 1125 (24%) were available somewhere on the web, and 76% were not.

Table 6. GPS: Electronic accessibility of citations

Yes	1125	24.1
No	3545	75.9
Total	4670	

The majority of electronic materials were journal articles.

Table 7. GPS: E-access by material type

Journal article	1098
Book chapter	10
Conference Proceedings	17
Total	1125

CONCLUSION

Based on this research, we can safely conclude that the great e-myth is just that. 38% campus-wide is hardly 'everything', and the situation is worse in the earth sciences. Except in journal publishing, the electronic revolution has barely started, and it's got a long way before it's even half-way successful. I would expect that biology and chemistry figures would be higher than the campus average, but we'll have to wait to see those analyses done by my colleagues in those areas.

In the meanwhile, I advise my fellow Geoscience Information Specialists to keep your bound journals, and to everyone, I would say: watch out what you're willing to believe.

Acknowledgements

I must acknowledge John McDonald, our Acquisitions Librarian, for teaching me the rudiments of Access, and everyone else on the Caltech Library Staff who designed and executed the project, of which this report is a very small part.

Appendix 1Campus: Dates of publication

2000	482
1999	183
1998	73
1997	38
1996	27
1995	17

1994	8
1993	4
1992	5
1991	1
1990	1
1988	1

1986	2
1984	3
1983	1
1982	2

Appendix 2 GPS: Top 20 titles cited

The top 20 journal titles cited by GPS Faculty (out of a total of 408) accounted for almost 54% of the citations. The "+" indicates title changes counted as one title. AGU publications accounted for 14.5% of all citations. GSA accounted for 5.7%. Nonprofit organizations accounted for 31% of the top 20, while commercial publishers produced 22%.

Title	**Times cited**	**% total citations**
+Journal of Geophysical Research	504	10.8
Earth & Planetary Science Letters	230	4.9
Science	221	4.7
Geochimica et Cosmochimica Acta	203	4.3
Nature	187	4.0
GSA Bulletin	129	2.8
Icarus	126	2.7
Geophysical Research Letters	122	2.6
Geology	99	2.1
Contributions to Mineralogy & Petrology	96	2.1
Lunar & Planetary Science Conference	83	1.8
+Meteoritics	78	1.7
Journal of Petrology	76	1.6
+Geophysical Journal	66	1.4
Bulletin of the Seismological Society of America	63	1.4
Tectonics	55	1.2
EOS	50	1.1
American Mineralogist	45	1.0
GSA Abstracts with Programs	38	0.8
Lunar & Planetary Institute Contributions	34	0.7
	2505	53.7