Big Data for Big Questions:
Assessing the Impact of Non-English Language Sources on Doctoral Research at Berkeley

*Susan Edwards, Lynn Jones, and Scott Paul McGinnis**

Introduction

In a 2014 guest editorial, Dan Hazen wrote "Research and teaching in such fields as public health, science, technology, and public policy today encompass international perspectives: the traditional area studies focus on language, culture, and history is no longer sufficient."[1] This statement raises important questions for academic libraries. In our increasingly globalized world, should research libraries extend support for foreign language acquisitions beyond the traditional strengths of area studies (traditionally history, language and culture) to other disciplines? Which disciplines use material published outside of the United States, and/or in languages other than English? How should research libraries make informed decisions about foreign language collecting policies?

Area studies librarianship has a long history at the University of California, Berkeley, with active collection development from East Asia, Middle East and North Africa, Russia and Eastern Europe, Latin America, sub-Saharan Africa, Southeast Asia, South Asia, and Western Europe representing about 30% of total selector funds. Collecting in the vernacular language is the standard practice, and since 1993, books have been added in 398 languages besides English, representing 39% of the total monographic acquisition, and 12% of the total circulation. Simple input and output measures like budget, percent of acquisitions, or even circulation are not, however, adequate indicators of the value this material adds to scholarship, or of research needs that may be unmet.

The Library supports doctoral research for about 100 departments at Berkeley, from African American studies to vision science (http://grad.berkeley.edu/programs/list/). Dissertations are an essential artifact of scholarship, and the sources they cite are a reliable indicator of the material students deem core to their research. Since Berkeley offers unlimited interlibrary loan at no charge to students, the sources cited are not restricted by what the Library owns.

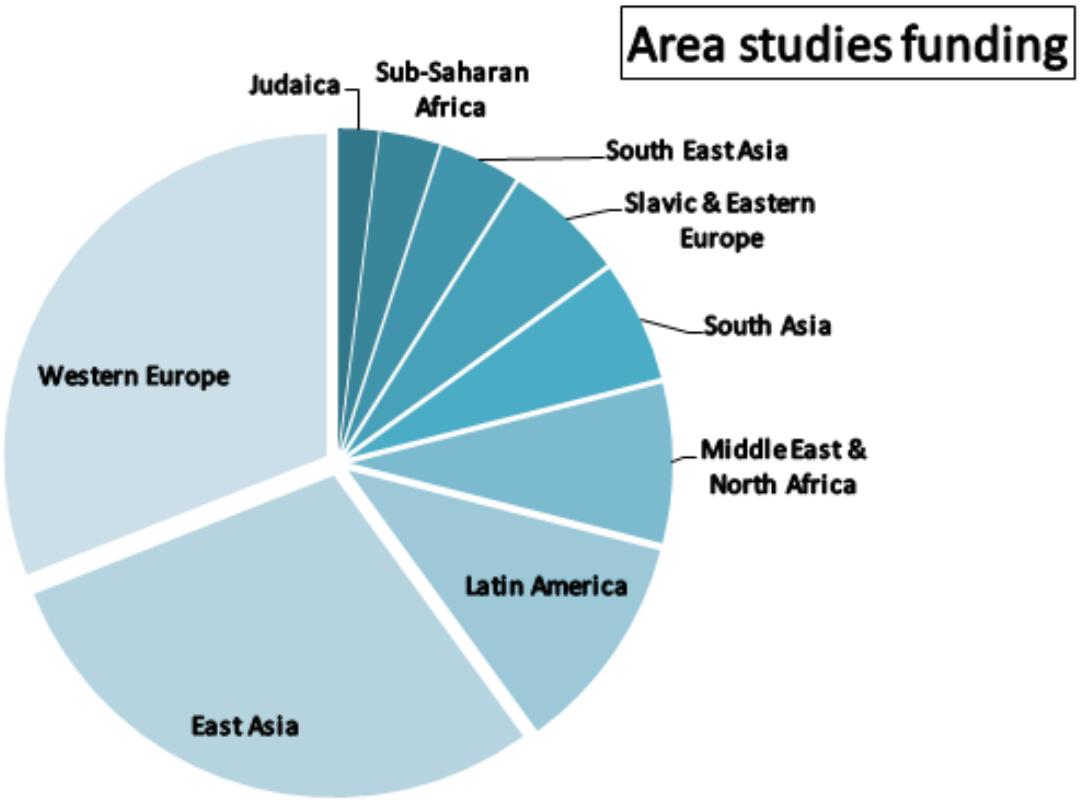
In a previous analysis of the ownership rate of material cited in dissertations in several disciplines at Berkeley, Edwards[2] discovered that language coding for cited references was included in the metadata available from ProQuest for dissertations in their Digital Dissertations database. This rich set of data made it possible to examine, in the present study, the languages of cited sources in dissertations from all disciplines at Berkeley from 2008-2015, and to cross-tabulate the language by format and department.

We found that non-English language sources were clustered in the traditional strengths of area studies: language, culture and history, including ten departments with a specific geographic focus, such as Slavic languages

* *Susan Edwards is Head, Social Sciences Division at The Library, University of California, Berkeley, seedwards@berkeley.edu; Lynn Jones is Reference Coordinator at Doe Library, University of California, Berkeley, lynnj@berkeley.edu; Scott Paul McGinnis is PhD candidate, History at University of California, Berkeley, spmgcinnis@berkeley.edu.*

and literatures, French, East Asian languages and literatures, as well as the subjects of history, religion, comparative literature and rhetoric. But we found some interesting variations between those disciplines, and some outliers from the social sciences. We confirmed what others have found: that disciplines with the highest percentage of non-English language usage also have the highest percentage of citations to monographs, rather than journals.

FIGURE 1
Area studies funding at Berkeley



Review of Literature

While there is a growing literature on citation analysis, we did not find any research that examined the language of the citations in dissertations from all disciplines. The following disciplinary-based articles were helpful in terms of how they defined the question, the methods they used for analysis, and the data they provided for benchmarking.

Knievel and Kellsey, in their article *Citation Analysis for Collection Development: A Comparative Study of Eight Humanities Fields,*[3] provided an unusually sophisticated analysis across eight humanities disciplines. Their study analyzed the language and format of 9,131 citations from the 2002 volumes of one journal in each of eight humanities fields: art, classics, history, linguistics, literature, music, philosophy, and religion. They found that while humanities overall cited 78% English, the variation between disciplines was substantial—from 65% in art to 99% in philosophy. French was the most commonly cited, followed by German and then Italian. They also analyzed format, and found philosophy and linguistics used fewer books (51% and 61%) compared with literature, music and religion (83%, 81% and 88%). This article was extremely valuable for providing benchmarks. Our findings confirmed these general patterns in the humanities versus the sciences or social sciences; and we also found varia-

tion between the disciplines within the larger subject groupings. Our research found that Spanish was the most frequently cited across all disciplines (after English), followed by French, German and Italian. These four languages comprised 75 percent of the non-English languages, followed by Japanese, Russian, Chinese, and Portuguese.

Schadl and Todeschini's *Cite Globally, Analyze Locally: Citation Analysis from a Local Latin American Studies Perspective*[4] analyzed 179 dissertations from 2000–2009 at the University of New Mexico on topics related to Latin America, and found that 85% of the citations were in English, 14% in Spanish and 1% Portuguese. For benchmarking, they used Kneivel and Kellsey's description of greater than 11.1% non-English as "extraordinary," 5–8% as dominant, and less than 2% as minimal. Schadl and Todeschini state, "It is important, however, to look carefully at graduate-level usage before making assumptions about value, especially in areas where well-established research strengths reflect unique institutional strengths."

Lenkart, Thacker, Teper, and Witt, in their article *Measuring and Sustaining the Impact of Area Studies Collections in a Research Library: Balancing the Eco-System to Manage Scarce Resources,*[5] analyzed five years of interlibrary lending transactions, focusing on University of Illinois, Urbana-Champaign's fulfillment of requests for non-English collections published in Less Commonly Taught (LCT) languages, and for materials with publication imprints located in regions outside the United States. We found the concept of LCT languages (anything other than German, Spanish and French), which has been used in the language teaching fields for decades,[6] to be helpful in our work. Lenkart et al., report two very interesting findings—that lending LCT material can have a significant national impact on scholarship, and that some research areas have a dominant LCT spoken language but publish a significant amount in English, French, German or Spanish. The latter fact influenced our decision to focus on language rather than country of publication for our research—though of course both language and country of publication are important in understanding the impact on scholarship of area studies collections.

Methodology

There is no single definition for the term 'big data'. According to Ward and Barker, "… big data is predominantly associated with two ideas: data storage and data analysis…. "Big" implies significance, complexity and challenge."[7] Bigness is, of course, relative. What is considered big now may be small in the future. What is big in one discipline may be small in another. For the purpose of this paper though, big data is a set of data too large for analysis using only MS Excel.

The data universe we studied was comprised of 938,000 citations from the bibliographies of 5668 dissertations from ninety-eight departments at UC Berkeley that conferred PhDs during the years 2008–2015. The data came from the ProQuest Digital Dissertations database (DDA), through a special arrangement with ProQuest, with additional information from Berkeley's Graduate Division.

Austin McLean, Director of Product Management at ProQuest, provided the references as .xml files, one per dissertation, along with a metadata file describing the dissertations and a schema file for validating the .xml. The citation elements were parsed into fields by ProQuest's proprietary algorithms, and included (in addition to the full text of the citation), publication format, title, article title, chapter title, subtitle, series title, authors' names, publisher name, place of publication, date of publication and language identification tag. Berkeley signed a license agreement with ProQuest regarding usage and sharing of the citation data.

Embargoed dissertations were not included in the ProQuest files. In general, Berkeley allows authors to embargo their dissertations for only two years, but ProQuest allows authors to embargo indefinitely. While all Berkeley dissertations are available through our catalog (including, after two years, those that were embargoed), they do not include the fielded citation metadata provided by ProQuest in DDA, and therefore are not able to be analyzed using the methods of this project. Thus,1,384 embargoed dissertations are not included in this project.

Manipulating the Data

We had no prior experience with such a huge dataset (almost one million citations, plus their metadata). Recognizing the limitations of Excel for this scale of data, the two librarian authors collaborated with the third author, a PhD candidate and digital humanities expert at Berkeley, on the analysis. ProQuest's metadata did not provide the department to which the dissertation was submitted, a key element needed for our analysis. Berkeley's Graduate Division provided another set of metadata, which included the departments but did not have a reference to ProQuest's dissertation number. An XQuery script merged the two into a single metadata file.

Certain analyses, including calculating how many citations overall were in non-English languages, cross-tabulations of book publishers by format, and publication formats cited in non-English languages, were performed on the whole dataset. Scott McGinnis programmed these queries, and provided Excel spreadsheets with subsets of the results, which were then analyzed using logical functions, pivot tables, and charting functions in Excel.

Once we had a unified metadata file able to associate citation records with departments, we could proceed with the analysis. XQuery scripts were written to mine the data and generate cross-tabulations as .csv files, which were then imported into Excel so that we could manipulate them. It is important to note that Excel, by its default behavior, has a troubling tendency to munge anything that looks like it could be a date (See Zeeberg et al.[8] and Ziemmann, Eren, and El-Osta[9] for the effect this long-standing problem has had on genetics research). Consequently, we had to import our data carefully, manually setting the encoding to UTF-8 and the fields to an alpha-numeric datatype (in Excel called simply "text"). This process added substantially to the time required to generate each report, but ensured that our page ranges would not be converted to dates (for example). Later, numerical text cells had to be reconverted to numbers for Excel mathematical functions to work.

Language Identification

As we started to work with the citations and language codes, we realized that some languages were misidentified. We found two types of errors. The first were sources identified as English, which weren't—for example, sources coded as English because the citation had been transliterated from a non-Roman script. While we found it quick and easy to manually identify whether or not a citation was English, we could not manually verify almost one million citations.

The second type of error incorrectly identified as a specific language citations which were not actually that language. Disambiguating related languages from small amounts of text—as in citations—is problematic for people not skilled in those languages, and is a known problem for computers as well. (See Vatanen, Jaako Vayrynen, & Virpioja[10]).The algorithm was more accurate in distinguishing English from non-English than identifying specific languages, but we were not sure how accurate. To verify the reliability of the coding of non-English by discipline we needed statistical expertise to develop a sampling strategy for manual verification. Fortunately, Berkeley has a free consulting service for staff and students, operated by doctoral students in statistics. (http://statistics.berkeley.edu/consulting).

The statisticians analyzed whether some languages were more inaccurately coded than others and determined that none were. They calculated sample sizes for each department with a significant percentage of non-English citations to achieve a 95% confidence level, wrote instructions for sampling, and created an Excel table for us to enter results. We then manually verified a random sample of citations (randomized with the Excel Random function) from those departments. The statisticians took the data we entered and calculated for each department the confidence interval and percentage (p value) for non-English citations.

Departments with a larger percentage of non-English language citations had better (smaller) confidence intervals. Departments that included fewer non-English citations had from moderate to large confidence intervals, making the estimates of non-English citations not very reliable.

Publisher Identification

Another initial goal of the research was to look for patterns in publishers of books cited by specific academic departments. This was expected to be a simple task, however, it proved impossible within our time, money and expertise constraints. A search of book citations yielded a list of over 55,500 unique entries in the publisher field. Visual analysis of this list revealed countless variations and errors in the citation of publisher names, which needed to be reconciled before a meaningful list of publishers could be produced. The most progress was made using OpenRefine, free software designed to "clean, transform, reconcile and match data," but this data defeated even OpenRefine. Twelve hours working with OpenRefine to reconcile a sample of variant publisher names yielded a list that still included too many errors to use. Future research in this area could benefit from using the WorldCat API to create an authority index of publisher names and variants.

TABLE 1
Departments with most dissertations, 2008-2015

Chemistry	419
Mechanical Engineering	322
Electrical Engineering and Computer Sciences	300
Molecular and Cell Biology	252
Physics	235
Civil and Environmental Engineering	229
Computer Science	202
Economics	181
Graduate School of Education	174
Environmental Science, Policy, and Management	171

Findings

The departments generating the most doctoral dissertations at Berkeley are in the sciences, with the exception of the Graduate School of Education.

The median number of citations per dissertation varies within disciplinary groupings. Disciplines were grouped according to where the academic department is located at Berkeley, so psychology and history are both included in the social sciences (see Appendix for list of departments and disciplinary groupings.)

FIGURE 2
Median number of citations per dissertation, by disciplinary group

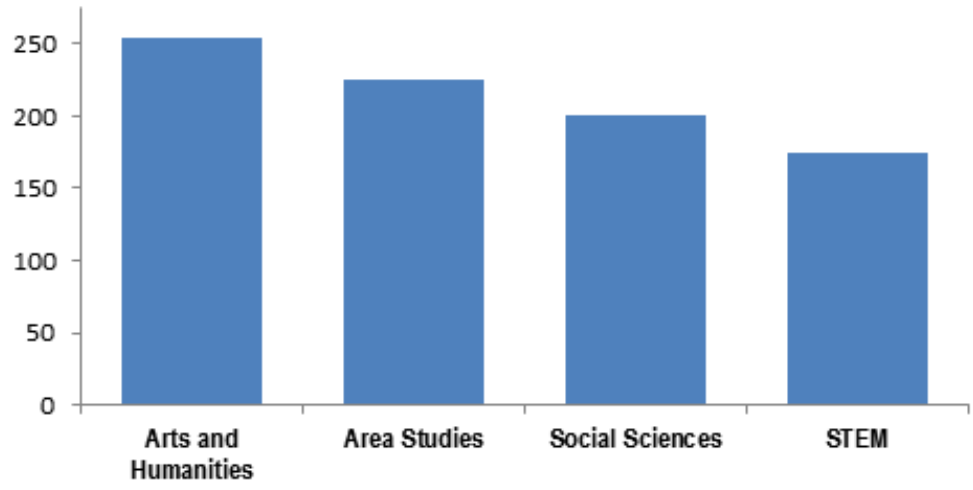


TABLE 2 Monographic-intensive departments
Departments citing more than 50% monographs
Asian Studies
Buddhist Studies
Chinese Language (part of East Asian Literature)
Classical Archaeology
Classics
Comparative Literature
English
Ethnic Studies
French
German
Hispanic Languages and Literatures
History
History of Art
Italian Studies
Jewish Studies
Latin American Studies
Medical Anthropology
Near Eastern Religions
Rhetoric
Romance Languages and Literatures
Slavic Languages and Literatures
South and Southeast Asian Studies

We found that the pattern of citation in the humanities still emphasizes monographs. Students in humanities departments cited more than 50% monographs in their bibliographies. We also observed substantial overlap between these departments and those citing a large percentage of non-English sources, lending credence to the common belief that most non-English language sources are monographs, rather than articles.

About 4% of all citations (41,211) were in non-English languages. The most frequently cited languages, in order, were Spanish, French, German, Italian, Japanese, Russian, Chinese, and Portuguese. The first four Western European languages comprised over 75% of the non-English citations.

FIGURE 3
Frequently cited languages

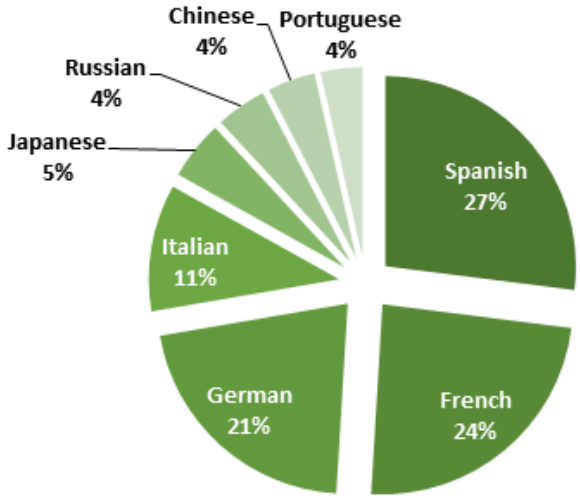
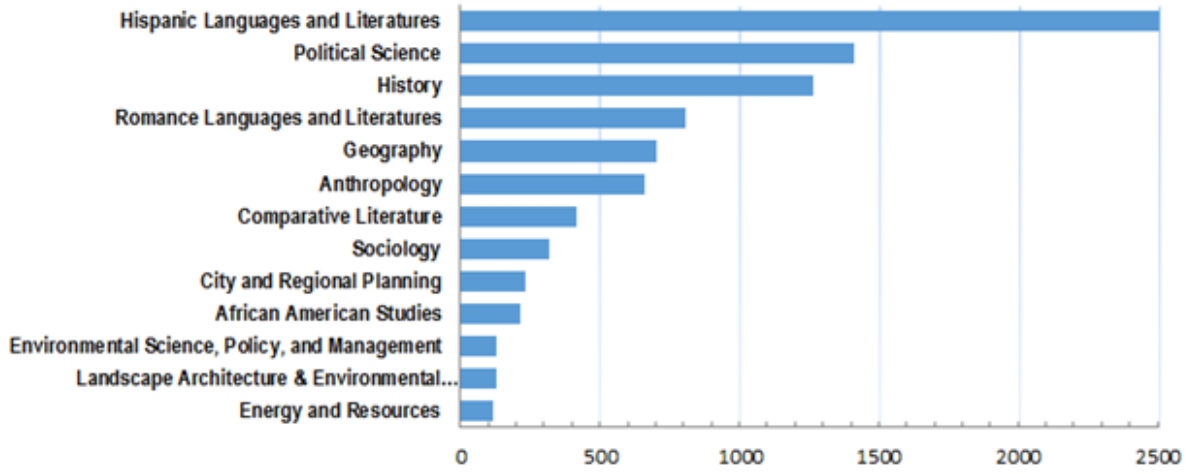


FIGURE 4
Departments frequently citing Spanish

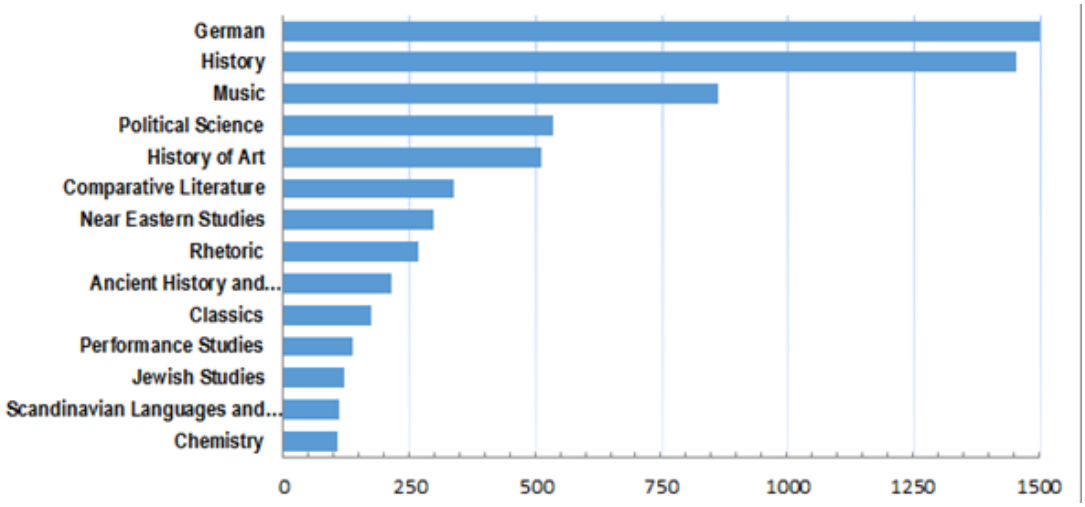


Language-intensive departments

Thirteen departments had more than 100 citations in Spanish. Latin American studies was just short of this number (97 citations).

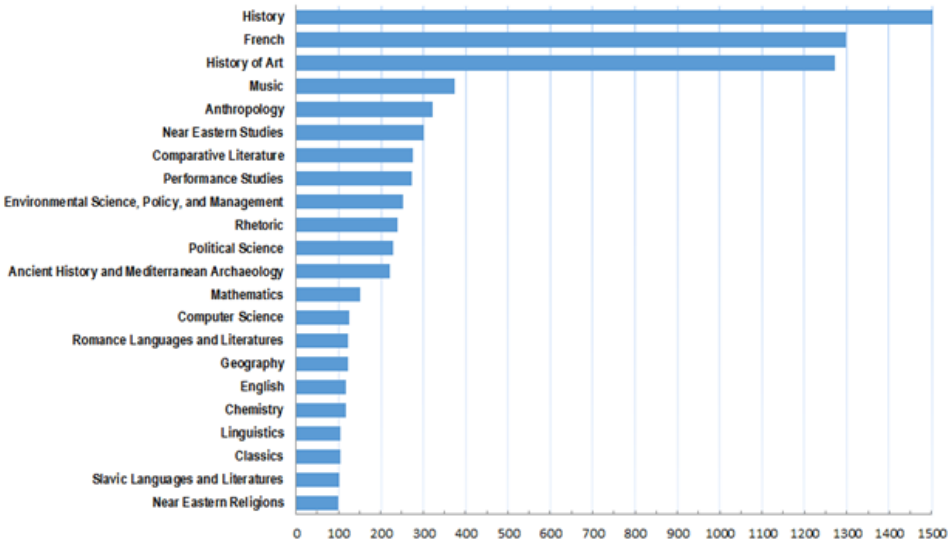
Fourteen departments had more than 100 citations in German, including chemistry, traditionally an area with strong German connections. Music notably cites a fair number of German sources.

FIGURE 5
Departments frequently citing German



Twenty-two departments cited more than 100 sources in French, a larger and more varied group (including math, chemistry and computer science) than for any other language. Interestingly, history cites French sources more than the French department does, because it produces many more dissertations. Art history cites French sources frequently, but there's a sharp drop-off in the numbers after that.

FIGURE 6
Departments frequently citing French



About a quarter of departments were responsible for 88% of all foreign language citations, including most of the LCT languages.

Departments with the highest percentage of non-English citations and with a large enough number of dissertations to have acceptable confidence levels.

TABLE 3
Departments with the highest percentage of non-English citations

Department	Non-English	Confidence Interval	# Dissertations	# Citations
Hispanic Languages and Literatures	57%	+/- 2%	32	6,284
French	55%	+/- 2%	14	2,916
Romance Languages and Literatures	53%	+/- 2%	14	3,273
Italian Studies	46%	+/- 2%	16	3,449
Music	45%	+/- 3%	30	7,483
German	45%	+/- 1%	22	5,275
Chinese Language (part of East Asian Languages and Literatures)	42%	+/- 1%	6	816
Slavic Languages and Literatures	42%	+/- 10%	18	4,239
Scandinavian Languages and Literatures	36%	+/- 14%	5	1,131
Jewish Studies	36%	+/- 9%	8	2,404
Buddhist Studies	34%	+/- 6%	3	1,509
History of Art	32%	+/- 2%	38	11,851
Ancient History and Mediterranean Archaeology	32%	+/- 6%	12	5,857
Comparative Literature	28%	+/- 5%	40	8,551
History	25%	+/- 2%	117	37,756
Near Eastern Studies	23%	+/- 4%	25	8,329
Geography	18%	+/- 6%	37	10,836
Rhetoric	12%	+/- 5%	37	8,902
Political Science	12%	+/- 3%	123	32,155
Anthropology	7%	+/-2%	87	27,261

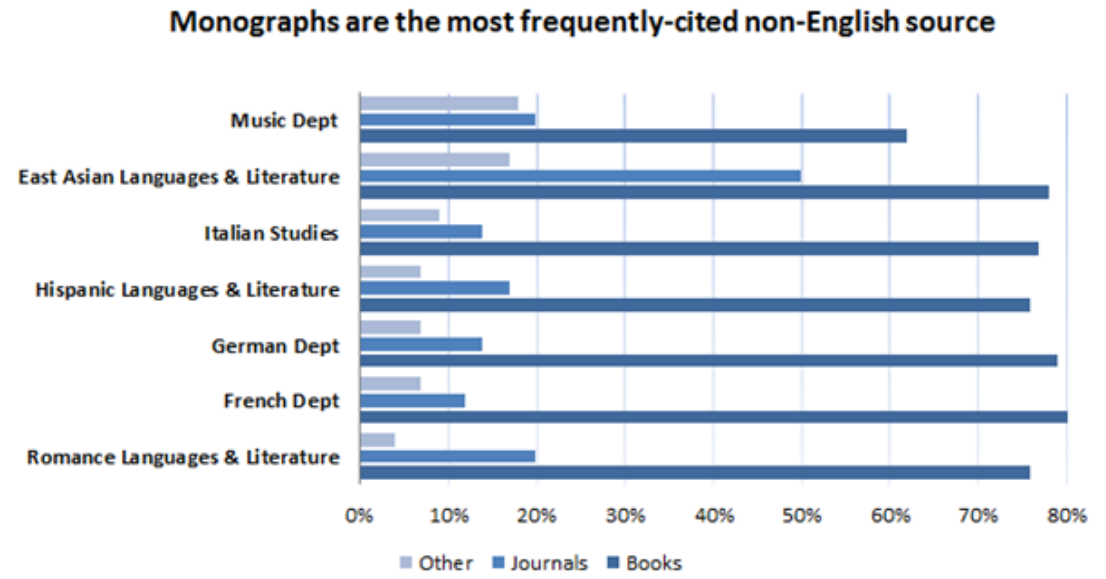
Monographs were most cited in the disciplines with the highest percentage of non-English citations

Monographs represent the greatest percentage of foreign language materials cited. The "other" materials cited are noteworthy. Music dissertations cite an unusually large percentage of "other," including 10% newspapers, probably for their reviews of performances. East Asian language and literature dissertations also cite a relatively large percentage of a variety of material types.

Languages Never Cited

Of the 398 languages Berkeley's library collects in 345 had no citations from any department on campus over the eight years of data that we examined. We cannot be certain, though, that any given language was never cited within the embargoed dissertations excluded from this study, and the unreliability of the language algorithm

FIGURE 7
Source types cited in other languages



leaves doubt about the accurate identification of LCT languages. Infrequently-cited language materials may also be used for Interlibrary Lending. The impact of infrequently cited languages warrants further investigation.

Discussion

Our research demonstrates that non-English language sources play a key role in scholarship at Berkeley. If we use Kneivel and Kellsey's benchmarks (greater than 11.1% non-English as "extraordinary"), then twenty departments at Berkeley used an extraordinary amount of non-English material in dissertations. An additional three departments used non-English material as a dominant component of their work.

In her 2015 textbook for area studies librarians, Pitman[11] pointed out the interdisciplinarity of area studies:

> "Understanding exactly what constitutes the research collection used by area studies researchers will always be difficult, because of the intrinsically interdisciplinary nature of the field. Although this is true of all interdisciplinary research, the growth of which has made collection boundaries more porous, libraries need to be careful to look out for overlaps between a carefully defined disciplinary collection, often under the management of a subject librarian, and a very much more diffuse collection on a country or region."

Our findings supported Pitman's call for greater collaboration between area studies librarians and other subject librarians. This was particularly true for Spanish, where the citation rate was significant in disciplines other than area studies, including political science, geography and anthropology.

We have known for years that non-English language monograph circulation is low, relative to English, and that the cost per circulation (not including cataloging) is high, but our research makes it clear that for about 25% of the departments at Berkeley non-English resources are essential for doctoral level research. It's also clear that many of these departments produce very few doctoral students, and that the seven departments with the largest number of doctoral students use virtually all English language material. This raises important philosophical and

values questions for collection development, outside the scope of this paper, but we hope that our data will help to inform the discussion.

Demonstrating that even in our global world, non-English language sources are little used by the sciences at Berkeley, our research also shows the flip side—that area studies disciplines at Berkeley rely heavily on English language sources. One of the questions we face is whether to collect in English, rather than, or in addition to, the vernacular language.

We confirmed that the LCT languages (those other than Spanish, French, German and Italian) were cited infrequently—in some cases, even by the department with which they are closely aligned. At a research institution there is constant pressure to add new geographic areas and vernacular languages to the curriculum, but our data on LCT languages shows that citation rate is likely to be low, even within the specific program, which may be a factor in the decision of whether to create new collections in LCT languages.

Conclusion

We began by referring to Hazen's thesis that our increasingly global world requires area studies to support more than its traditional areas. We were not able to find significant use of non-English materials in the STEM and public policy areas at Berkeley, though we did see non-English language use in social science disciplines. It is possible that the international perspective Hazen reported does exist, but that English so thoroughly dominates academic writing (even from other countries) that non-English language citation wasn't necessary. Further research into the place of publication of materials cited in non-area studies fields might provide another perspective on this question.

The study confirmed that four Western European languages (Spanish, French, German and Italian) comprise most of the non-English languages cited, even at a major research university with extensive Less Commonly Taught language collections. This does not suggest abandoning the collection of materials in the world's languages: quantity of research is not the only measurement of importance. Closer study of citations of lesser-cited languages would be useful as a way to understand what research is enabled by these materials. Furthermore, dissertation citations are only one measure of use. The results should be corroborated by interlibrary loan usage and research into faculty citations.

Besides our goal to learn about citation of non-English language materials, the project also tested "big data" analysis methods for bibliometrics. We are convinced that these methods can extract valuable information from dissertation citations for use by selectors and collection managers. However, the difficulties we encountered will require librarians to develop (or hire) expertise in programming, machine learning, and data management to pursue this research.

Acknowledgements

The authors would like to thank the following, without whom this research would not have been possible: Austin McLean, for his assistance in providing ProQuest's metadata for Berkeley's dissertations, the Librarians Association of the University of California, for a research grant which supported Scott's time, and Yuansi Chen and Simon Walter of Berkeley's Statistical Consulting Service, doctoral students in statistics, who made it possible for us to have confidence in the statistical validity of our results.

Appendix 1. Disciplinary groups at Berkeley, 2016

Arts and Humanities
Ancient History and Mediterranean Archaeology
Classical Archaeology
Classics
Comparative Literature
English
Film and Media
French
History of Art
Music
Performance Studies
Philosophy
Rhetoric

Area Studies
Asian Studies
Buddhist Studies
Chinese Language
German
Hispanic Languages and Literatures
Italian Studies
Japanese Language
Jewish Studies
Latin American Studies
Near Eastern Religions
Near Eastern Studies
Romance Languages and Literatures
Scandinavian Languages and Literatures
Slavic Languages and Literatures
South and Southeast Asian Studies

Social Sciences
Anthropology
Architecture
Business Administration
City and Regional Planning
Demography
Economics
Education
Educational Leadership
Environmental Planning
Ethnic Studies
Geography
History
Information Management and Systems
Jurisprudence and Social Policy
Landscape Architecture & Environmental Planning
Linguistics
Medical Anthropology
Political Science
Psychology
Public Policy
Science and Mathematics Education
Social Welfare
Sociology
Sociology and Demography
Special Education

STEM
Agricultural and Environmental Chemistry
Applied Mathematics
Applied Science and Technology
Astrophysics
Bioengineering
Biophysics
Biostatistics
Chemical Engineering
Chemistry
Civil and Environmental Engineering
Comparative Biochemistry
Computer Science
Earth and Planetary Science
Electrical Engineering and Computer Sciences
Endocrinology
Energy and Resources
Environmental Health Sciences
Environmental Science, Policy, and Management
Epidemiology
Health Policy
Industrial Engineering and Operations Research
Infectious Diseases and Immunity
Integrative Biology
Logic and the Methodology of Science
Materials Science and Engineering
Materials Science and Mineral Engineering
Mathematics
Mechanical Engineering
Metabolic Biology
Microbiology
Molecular & Biochemical Nutrition
Molecular Toxicology
Molecular and Cell Biology
Neuroscience
Nuclear Engineering
Physics
Plant Biology
Public Health
Statistics
Vision Science

Appendix 2. Languages cited [not verified]

Afrikaans	44
Albanian	6
Arabic	728
Asturian; Leonese	1
Basque	1
Breton	1
Chamorro	1
Chinese	1,495
Croatian	12
Czech	313
Danish	46
Dutch	370
English	1
Esperanto	2
Finnish	41
French	8,837
German	7,872
Hebrew	226
Hindi	233
Hungarian	1
Icelandic	350
Indonesian	296
Irish	1
Italian	3,990
Japanese	1,849

Korean	198
Kyrghyz	20
Latin	3
Latvian	85
Lithuanian	16
Malay	54
Modern Greek	9
Mongolian	7
Norwegian	90
Persian	2
Polish	226
Portuguese	1,304
Romanian	33
Russian	1,583
Scottish Gaelic	1
Slovenian	15
Spanish	9,959
Swedish	61
Swiss	7
Thai	5
Tibetan	11
Turkish	494
Ukrainian	93
Vietnamese	111

Appendix 3. Departments sampled for non-English language citations

African American Studies
Ancient History and Mediterranean Archaeology
Anthropology
Buddhist Studies
Chinese Language [part of East Asian Language and Literature]
Classical Archaeology
Comparative Literature
Film and Media
French
Geography
German
Hispanic Languages and Literatures
History
History of Art

Italian Studies
Japanese Language
Jewish Studies
Landscape Architecture & Environmental Planning
Latin American Studies
Music
Near Eastern Studies
Performance Studies
Political Science
Rhetoric
Romance Languages and Literatures
Scandinavian Languages and Literatures
Slavic Languages and Literatures
South and Southeast Asian Studies

Notes

1. Dan Hazen, "Researching Library Support for International Studies: Successes to Celebrate, Goal Posts to Move," *College & Research Libraries* 75, no. 4 (July 1, 2014): 418–21, doi:10.5860/crl.75.4.418.
2. Susan E. Edwards, "Making Hard Choices: Using Data to Make Collections Decisions," *6th International Conference on Qualitative and Quantitative Methods in Libraries (QQML 2014)*, May 29, 2014, http://escholarship.org/uc/item/3429v849.
3. Jennifer E. Knievel and Charlene Kellsey, "Citation Analysis for Collection Development: A Comparative Study of Eight Humanities Fields," *The Library Quarterly* 75, no. 2 (April 1, 2005): 142–68, doi:10.1086/431331.
4. Suzanne M. Schadl and Marina Todeschini, "Cite Globally, Analyze Locally: Citation Analysis from a Local Latin American Studies Perspective," *College & Research Libraries* 76, no. 2 (March 1, 2015): 136–49, doi:10.5860/crl.76.2.136.
5. Joe Lenkart et al., "Measuring and Sustaining the Impact of Area Studies Collections in a Research Library: Balancing the Eco-System to Manage Scarce Resources," 2015, https://www.ideals.illinois.edu/handle/2142/73824.
6. "Report of the MLA Task Force on the Less Commonly Taught Languages:," *Foreign Language Annals* 11, no. 6 (1978): 641–45.
7. Jonathan Stuart Ward and Adam Barker, "Undefined by Data: A Survey of Big Data Definitions," *arXiv Preprint arXiv:1309.5821*, 2013, https://arxiv.org/abs/1309.5821.
8. Barry R Zeeberg et al., "Mistaken Identifiers: Gene Name Errors Can Be Introduced Inadvertently When Using Excel in Bioinformatics," *BMC Bioinformatics* 20045, no. 80 (June 23, 2004), doi:10.1186/1471-2105-5-80.
9. Mark Ziemann, Yotan Eren, and Assam El-Osta, "Gene Name Errors Are Widespread in the Scientific Literature," *Genome Biology* 201617, no. 177 (August 23, 2016), doi:10.1186/s13059-016-1044-7.
10. Tommi Vatanen, Jaakko J Vayrynen, and Sami Virpioja, "Language Identification of Short Text Segments with N-Gram Models," vol. 10, 2010, http://www.lrec-conf.org/proceedings/lrec2010/pdf/279_Paper.pdf.
11. Lesley Pitman, *Supporting Research in Area Studies: A Guide for Academic Libraries* (Amsterdam, Boston: Chandos Publishing, 2015).