Executive Summary

From February 19th to March 2nd 2007, Google performed a survey of books across the collections of the library of the Universidad Complutense de Madrid (UCM). Following are the high-level results of this survey.

Population Size	Sampling %	% Scannable	% Unscannable	Unscannable Reason	
				% Oversize	% Poor Condition
123138	3.30%	73.43%	26.57%	10.23%	18.54%

Based on our survey, we anticipate being able to scan about **99,000**, or 73%, of the roughly 135,000 books contemplated under the Google-UCM digitization agreement.

Purpose of Survey

Google conducted the collection survey in order to determine the number of books held by the UCM libraries that would be suitable for digitization as part of the Google Book Search Library Project. The findings of the survey will help us accurately plan the appropriate scanning capacity for this project, as well as the expected length of time it will take for the project to be completed.

Methodology

To perform the survey, Google hired a temporary labor force consisting of 8 staff and 2 supervisors. Surveyors were given a pre-generated "pick list" of library call numbers, and instructed to pull books on the list from library shelves and inspect them according to the following factors:

- Publication Date
- Size
- Condition
- Binding Type

Not all of the books in each library were cataloged. In order to avoid biasing the survey in favor of cataloged books, surveyors were also instructed to pull and inspect a certain number of books at random from the shelves of each library, typically by counting books on shelves and inspecting every *n*th book, where *n* was based on the number of uncataloged books in the library to be evaluated.

Google generated the list of call numbers from a file of pre-1867 holdings records
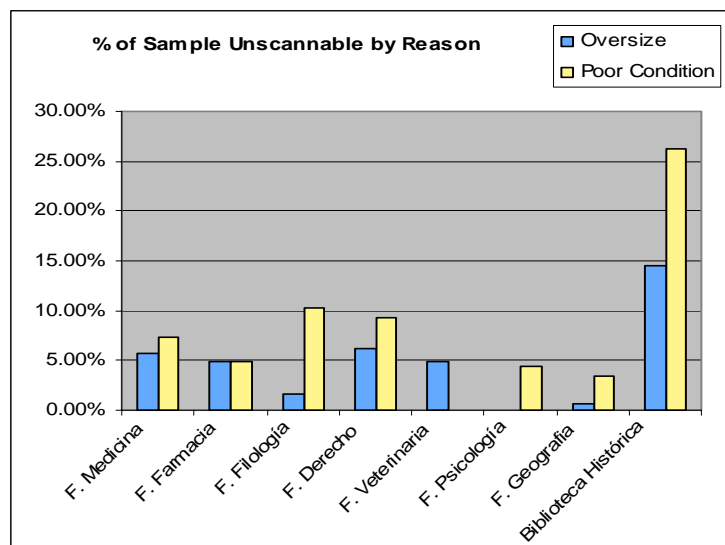
provided to Google by the UCM. 8 libraries were covered in the survey. UCM estimated that together these libraries hold 123,198 volumes with publication dates before 1867. Based on this population, Google decided on a sample size of 4137, or 3.36%. Of this number, 440 books surveyed fell outside of the pre-1867 range, making the actual sample size for the pre-1967 population 3,697, or 3%. 56% of this sample came from the generated picklists; the remaining 44% came from random pulls from library shelves.

Aggregated Results

Facility	Population Size	Sampling %	% Scannable	% Unscannable	Unscannable Reason	
					% Oversize	% Poor Condition
F. Medicina	10014	3.62%	88.42%	11.58%	5.65%	7.34%
F. Farmacia	2608	4.83%	90.16%	9.84%	4.92%	4.92%
F. Filología	8188	3.96%	88.40%	11.60%	1.57%	10.34%
F. Derecho	10460	3.94%	85.71%	14.29%	6.27%	9.27%
F. Veterinaria	2417	3.48%	95.18%	4.82%	4.82%	0.00%
F. Psicología	1798	2.56%	95.65%	4.35%	0.00%	4.35%
F. Geografia	8726	3.30%	95.83%	4.17%	0.69%	3.47%
Biblioteca Histórica	78927	3.06%	62.50%	37.50%	14.49%	26.20%

Scannable/Unscannable books

Google will scan all books that meet its scanning criteria. Books deemed unscannable fall into two (possibly overlapping) categories: Oversized and Poor Condition. Oversized books are those that are too large to fit our scanning technology. Poor condition books are those books whose physical condition would compromise the quality/success of the scan, and/or result in risk of significant damage to the book during the scanning process. Below is a graph showing unscannable books at each library, coded by reason. Note that some books may be both oversized and in poor condition.

Poor condition books
When identifying a book as Poor Condition, we look for the following factors:

- Text block damage
- Inner joint damage
- Torn pages
- Brittle pages

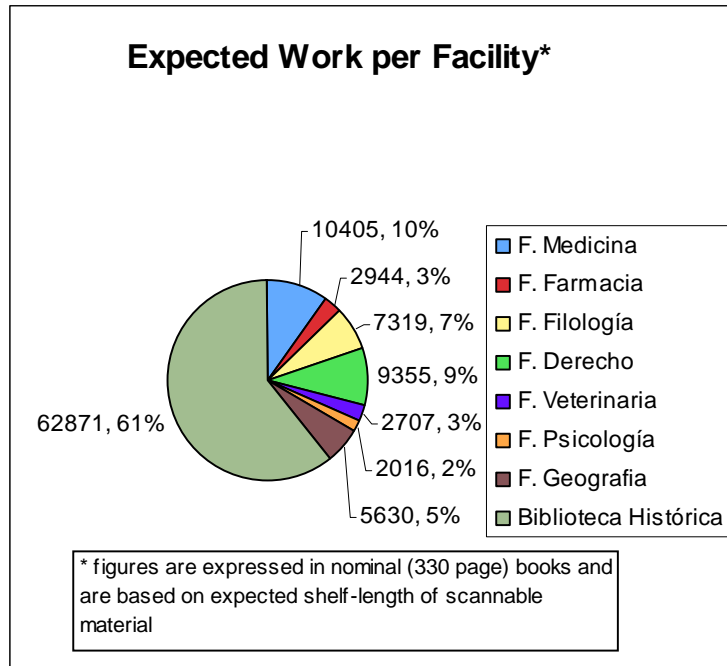Below is a table indicating the breakdown of Poor Condition factors in books from each of the surveyed libraries.

Facility	% Text Block	% Inner Joint	% Torn Pages	% Brittle Pages
F. Medicina	5.37%	4.80%	0.85%	1.13%
F. Farmacia	1.64%	1.64%	0.82%	0.82%
F. Filología	8.78%	2.82%	1.25%	0.63%
F. Derecho	3.76%	2.76%	1.25%	0.75%
F. Veterinaria	0.00%	0.00%	0.00%	0.00%
F. Psicología	4.35%	4.35%	4.35%	4.35%
F. Geografia	0.00%	0.00%	0.00%	0.00%
Biblioteca Histórica	10.67%	16.65%	3.49%	5.07%

Not found books
Of the 2,318 books on the surveyors' picklist, surveyors were unable to find 54. A list of these books is included in the accompanying spreadsheet.

Expected contributions per library
As the chart below indicates, the lion's share of books for the project are anticipated to come from the Biblioteca Historica. This library also contains the largest number and percentage of unscannable books.

Expected Work per Facility*

10405, 10%

2944, 3%

7319, 7%

9355, 9%

2707, 3%

2016, 2%

62871, 61%

5630, 5%

- ■ F. Medicina
- ■ F. Farmacia
- □ F. Filología
- ■ F. Derecho
- ■ F. Veterinaria
- ■ F. Psicología
- ■ F. Geografia
- ■ Biblioteca Histórica

* figures are expressed in nominal (330 page) books and are based on expected shelf-length of scannable material

Next steps

Based on these results, Google will work with UCM to develop a plan to approach the transport of books from these libraries for digitization, and an expected timeframe for the project.