Patricia A. Soler
Ph.D. candidate
Latin American Literature and Cultural Studies
Georgetown University
Research focus: Spanish and Portuguese Literary Works

SUMMARY

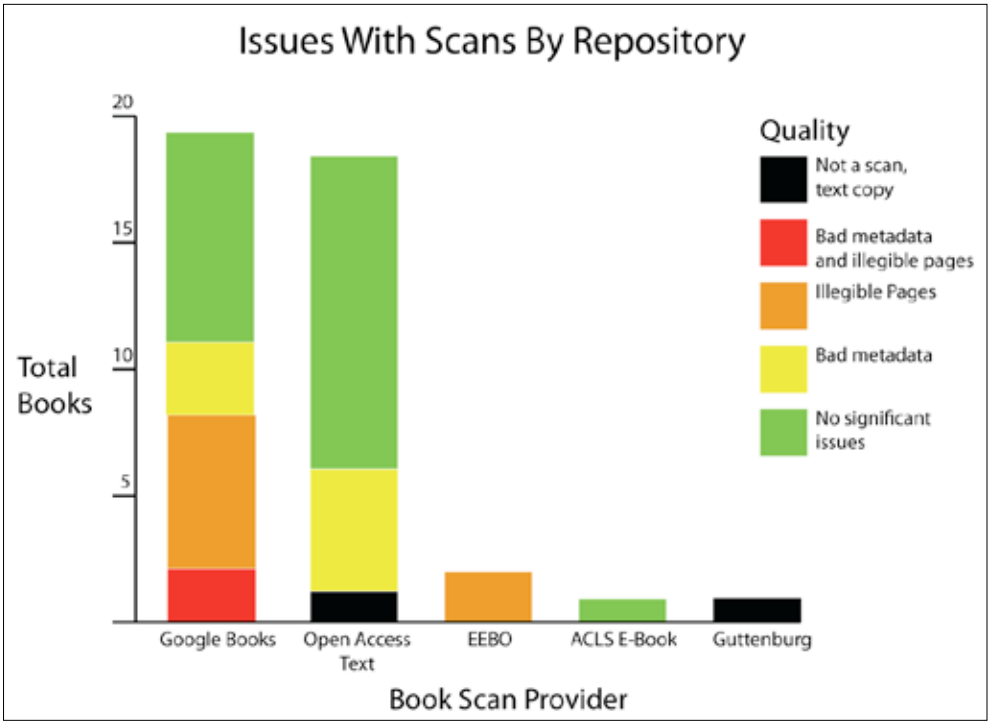
Scope

This document distills and summarizes the results of Patricia Soler's research on the quality and availability of digitized Spanish and Portuguese literary works, both in their original languages and in English translation. Her research offers a particularly valuable comparison of digitizations in Google Book Search (GBS) and the Internet Archive's Open Access Text Archive (OATA).

Overview

Over the summer of 2008, Soler analyzed digitized copies of 41 Spanish and Portuguese literary works, both in their original languages and in English translation. Her searches included a diverse set of books published between 1688 and 1964. Of the 41 scans examined, 19 were from GBS, 18 from OATA, 1 from ACLS Humanities E-Book, 2 from Early English Book's Online (EEBO) and 1 from Project Gutenberg. An overview of the quality of scans found in each repository is presented in the figure below.

Digitizations marked as having *illegible pages* had between 1 and 15 illegible pages. Digitizations marked as having *bad metadata* had



significant problems with core metadata, e.g., wrong author name, wrong publisher, or significantly wrong dates of publication. Two of the digitizations Soler worked with were not scans at all, but were text-only copies of books.

Key Findings

Illegibility was highest in GBS and EEBO. Of the 41 books Soler analyzed, 10 (24%) had illegible pages. All of the books with illegible pages came from either GBS or EEBO. Both of the books Soler analyzed from EEBO had illegible pages—in one case three illegible pages, in the other four. In GBS, 8 of the 19 books, or 42% of the books Soler analyzed from GBS, had illegible pages. Most of the GBS scans with illegible pages had only one or two problematic pages. However, one Google scan had 15 illegible pages.

87% of scans were identified as very high quality. Soler ranked all but four of the scanned works as either excellent or above average. Of the four scans she marked as poor quality, three came from GBS and one from EEBO. While there were significant issues with metadata on some of these works, the overall quality of the scans, particularly in the OATA, was quite high.

60% of works had minor search issues. Soler did not uncover any major issues with the quality of search results in the scans. She found that 16 of the digitizations had effectively error-free OCR, while 25 had minor search issues. Of the 25 digitizations with minor search issues, most involved problems with detecting accented characters (á, é, í, ó, ú, ñ) in Spanish and Portuguese works.

Some data scans were particularly problematic. As indicated in the figure on the previous page, OATA's scans analyzed were of higher quality than those from GBS. With that noted, there were still several issues that were particularly problematic with OATA's offerings. One book was bound with an unrelated book, one had almost no core metadata, and several had incorrect dates of publication. Soler recognized that the scans of particularly high quality had been originally scanned by Microsoft's Live Book Search project. In one case, when looking for translations of the writings of Christopher Columbus, OATA returned 20 digitizations of the same text, many of which were duplicate scans of the same edition, and many of which had incorrect data about their date of publication, authors, and translators. In this case, Soler needed to click through to see what exactly had been scanned in each case. In one other situation, OATA returned a text-only copy of a book without any page images. This text copy came with no metadata about the work and no information about how, or why, the plain text copy was created.