Lynn Silipigni Connaway
Timothy J. Dickey
OCLC Research

Publisher Names in Bibliographic Data: An Experimental Authority File and a Prototype Application

Note: This is a pre-print version of a paper published in *Library Resources and Technical Services* Please cite the published version; a suggested citation appears below. Correspondence about the article may be sent to lynn_connaway@oclc.org.

Abstract

The cataloging community has long acknowledged the value of investing in authority control; as bibliographic systems become more global, the need for authority control becomes even more pressing. The publisher description area of the catalog record is notoriously difficult to control, yet often necessary for collection analysis and development. The research presented in this paper details a project to build a database of authorized names for major publishers worldwide. ISBN prefix data were used to cluster bibliographic records based on publishing entities; the resulting database contains thousands of variant forms of each publisher's name, and data about their overall publishing output. Profiles of four large publishers were compared: each publisher's languages of publication, formats, and subjects demonstrated their distinctive publishing output, and validated the record clusters. Finally, the results of the research were made freely available on the Web via a prototype set of web pages displaying the publishing profiles of more than eighteen hundred major publishers.

© 2011 OCLC Online Computer Library, Inc.
6565 Kilgour Place, Dublin, Ohio 43017-3395 USA
http://www.oclc.org/

Reuse of this document is permitted consistent with the terms of the Creative Commons Attribution-Noncommercial-Share Alike 3.0 (USA) license (CC-BY-NC-SA): http://creativecommons.org/licenses/by-nc-sa/3.0/.

Suggested citation:
Connaway, Lynn Silipigni, and Timothy J. Dickey. 2011. “Publisher Names in Bibliographic Data: An Experimental Authority File and a Prototype Application.” *Library Resources and Technical Services*, 55,4. Pre-print available online at: http://www.oclc.org/research/publications/library/2011/connaway-lrts.pdf

Acknowledgements

The authors would like to thank Jeremy Browning, Clifton Snyder, and Erin Hood, OCLC Research, and Akeisha Heard, formerly of OCLC Research for their contributions to this research.

Note

This research was conducted when Timothy J. Dickey was a post-doctoral researcher at OCLC Research, Dublin, Ohio. He currently is teaching in the library science programs of Drexel University, Kent State University, and San Jose State University.

> "The centrality of authority control in librarianship and its value to the user is not likely to change soon." –Nirmala Bangalore and Chandra Prabha, 1998.[i]

Introduction and Research Goals

A 1979 international library technology conference dubbed authority control, defined as the creation and maintenance of standardized links between the various forms of an access point, "The Key to Tomorrow's Catalog." [ii] Despite dissenting views that authority files would be prohibitively difficult and expensive, the conference attendees believed that such files would give structure to the burgeoning universe of knowledge, fulfilling the objectives of Charles Cutter for the 21st century. In the decades since, the library community has slowly but surely progressed towards the goal of universal authority control; local electronic authority files proliferated, followed by larger collaborative efforts such as the Name Authority Cooperative (NACO) (www.loc.gov/catdir/pcc/naco), led by the Library of Congress, and the Virtual International Authority File (VIAF) (viaf.oclc.org), hosted by OCLC. Yet among all of the data elements in MARC cataloging that could benefit from authority control, the publisher description area – and specifically publisher names – have no authorized forms.

The goal of the research reported here is to develop a service to support advanced collection analysis and publisher entity and user discovery services. Specifically, it is a project to cluster items in library collections based upon the entity that published or distributed them. The objectives of the research are:

I. To build a database that will

A. Identify:

- Authoritative strings for publishers
    - Common variants of the preferred/ authoritative version of the name
    - Common variants for the locations of publishers
- Hierarchical references to variants and related entities and nesting of subsidiaries
- Definitions of publishing entities
    - Data-mined information regarding formats, languages, subjects, etc. for each entity

B. Conform to international authority and standards practice.

II. To develop a method to:

A. Integrate the mapping of the database entries to WorldCat bibliographic records

B. Automate updates of the publisher data

This paper reports the results of the first stages of the project, the building of a publisher name authority database and the development of a prototype web interface with the bibliographic records associated with each publisher in the database.

Researchers explored a number of different technologies and methods for the clustering of bibliographic records. These clusters were ultimately constructed on the basis of metadata relating to the issuing entities, specifically metadata in the Publisher Description Area (MARC field 260) and in International Standard Book Numbers (ISBNs, MARC field 020). Along the way, the aggregate of the records that could be assigned to different publishing entities allowed researchers to gain intelligence about the nature of individual publishers, producing rich portraits of their global presence and publication patterns. This intelligence, achieved through data mining and through broader research, can be valuable for libraries' collection intelligence (both collection analysis, and intelligence related to approval plans and acquisition patterns); in addition, the data collected about individual publishers has value for both librarians and publishers related to overall subject coverage, and "family trees" among publishers and their various imprints, subsidiaries, and acquisitions.

The results were twofold: an experimental Publisher Name Authority File and a prototype set of web pages that expose the various data about each publisher and its publication footprint. The database of publishers includes more than eighteen hundred high-incidence publishers, with operations in fifty-seven countries worldwide. A total of more than sixty thousand variants have been mapped onto the preferred form of each publisher's name, resulting in distinct bibliographic profiles comprising some 16.3 million records in total. All of the data for each publishing entity are freely viewable via the WorldCat Publisher Pages (http://worldcatpubs.oclc.org/wcp), including the complete organizational chart for each complex of publishers.

Literature Review

At the library technology conference referenced above, despite dissenting views that authority control would be prohibitively difficult and expensive, the conference attendees believed that if properly controlled, such files would give structure to the bibliographic universe and the universe of knowledge.[iii] One well-known definition of authority control is "the process of maintaining consistency in the verbal form used to represent an access point in a catalog and the further process of showing the relationships among names, works, and subjects."[iv] The practical (if anecdotal) experience of librarians did lead to research into the high cost of authority files. The proliferation and popularity of local authority files have increased the breadth of authority control over the names of both individuals and corporate bodies. A special issue of *Cataloging & Classification Quarterly* followed the international conference "Authority Control: Definitions and International Experiences" (Florence, IT, Feb. 10-12, 2003). [v] Various projects reported there included local authority files for historical corporate bodies in the Bibliothèque Nationale de

France, corporate and personal names associated with the worldwide Roman Catholic Church, experiments in interoperability among disparate Italian authority files and among CJK catalog standards, as well as supportive theoretical arguments in favor of the practice of authority control.

Several studies have tested automatic solutions to creating authority files, with mixed results. Snyman and Rosenberg first addressed the need to develop new technological and automatic solutions to the cost of maintaining authority control.[vi] Veve reported on a project at the University of Tennessee Library's conclusion that despite various efforts to automate authority work, levels of human intervention were still required, though perhaps the costs of this effort could be held down.[vii] Patton et al. attempted to explore automated processes to assist catalogers in name authority control, by automatically calculating the probability of matches between metadata strings and Library of Congress authority files.[viii] Their matching algorithm was successful 58% of the time. Rodriguez, Bollen, and Van de Sompel explored a more general solution to propagate metadata from environments rich in metadata onto resources whose metadata are sparse.[ix] Specifically dealing with corporate name authorities, Blake and Samples reported on a project at North Carolina State University to normalize organization names within the Libraries' electronic resources management (ERM) module that seeks the benefits of greater data integrity in their management of vendors and acquisitions.[x]

Progress has been made, as well, in the internationalization and aggregation of name authority control. The Name Authority Cooperative (NACO) was founded in 1976 as a library consortium that, under the leadership of the Library of Congress, maintains an extensive name authority database; Bynum offers a prospectus of the history and operation of the organization.[xi] The presence of corporate names within the NACO authority file has materially aided the construction of the OCLC Publisher Name Authority File (PNAF). More recently, OCLC Research has led and hosted the construction of the VIAF. After a prototype was launched that virtually combined the national name authority files of the Library of Congress, the Deutsche

Nationalbibliothek, and the Bibliothèque Nationale de France, the VIAF has grown into a virtual collaboration among eighteen national-level cataloging organizations.

Research at the University of Virginia attempted to use mathematical techniques of clustering theory to aggregate different authority files; French, Powell and Shulman worked with Astrophysics Data System records, using a subset of the database comprising 85,000 refereed articles from seven different journals.[xii] The researchers admit at the outset that it is completely impossible for them to completely automate the process of clustering corporate names "by lexical techniques alone."[xiii] Instead, they used an iterative variety of programmatic techniques for string clustering and matching, and approximate word matching, followed by expert review of the results; their complete technique achieved cost savings of approximately half of the human effort in constructing an authority file. Similar clustering techniques also featured in the early stages of the PNAF (see below).[xiv]

Authority control over the names of publishers, however, in current cataloging practice continues to present some difficult issues. The MARC field 260, subfield b ($b), contains the name of the publisher in "the shortest form of the name that it can take to be understood internationally," with the added complexity that local practices may stipulate how much text to transcribe exactly from the title page or other chief source of information.[xv] In addition, changes in the rules of cataloging practice compound the difficulty of automatically identifying matches in the data strings found in this part of the record (see table 1). A study by Jin found a discrepancy rate of 25 percent between corporate names as the organizations called themselves on official company websites, and the corporate names in the Library of Congress authority file.[xvi] However, collection development in libraries often depends on the specialized nature of a publisher's output.[xvii] The ALCTS has reported on the need for better authority control for library acquisitions.[xviii]

Methodology

The work by OCLC Research to normalize publisher names has involved the methodologies of data mining and programmatic clustering of bibliographic records, supplemented by manual review of the results. Data mining appeared first as a tool for business intelligence, only later to be adopted by libraries; the success of Google and Amazon has taught the library field that greater value exists within bibliographic data as well. Libraries have made huge investments in creating and maintaining rich, structured information describing the resources in their collections. These data embody considerable value by supporting basic local access and inventory control. They also represent potential value in terms of knowing more about the characteristics of library collections. OCLC's Office of Research has invested significant effort in the area of data mining.[xix]

Specifically, research projects have demonstrated the value of the WorldCat database as an "aggregate collection" of bibliographic data.[xx] It thus has utility as a global-scale dataset of potential value that can "not only provide librarians data for decision-making for collection and service development, but also provide users with enhanced discovery and access methods."[xxi] The WorldCat database is an increasingly global and increasingly comprehensive source of bibliographic data, and remains strongest in its data on books. As of February 2011, WorldCat contains more than 217 million records, with more than 1.68 billion distinct library holdings of those resources; some 57.5 percent are non-English catalog records, illustrating the increasingly global reach of the "aggregate collection."[xxii] Its member libraries are located in more than one hundred countries, and the data go beyond those countries to include works from those countries that are collected in other OCLC member libraries.

The first OCLC research into publisher name data was performed on an earlier snapshot of the WorldCat database from July 2005. Researchers mined bibliographic records that had a value of English in the language fixed field, resulting in 35,434,911 records (61 percent of the database). The next criterion was to explore the presence of valid ISBNs in the 020 field, as ISBN prefixes provided a more consistent way of identifying publishers. Researchers determined that in 2006 more than 22 percent of WorldCat records contained an ISBN, and more than 99 percent of those ISBNs contained a valid check digit (The check digit in an ISBN is the final digit of the number; it is not assigned by the publisher, but rather is computed according to rules set out by the International ISBN Authority, and serves as a validation for ISBN data points..[xxiii]

Having demonstrated the prevalence of ISBN data within the bibliographic records, researchers made a first experimental attempt at programmatic clustering of records with a single ISBN prefix, to gather variant forms of publisher names. Data clustering is defined by *The Free Dictionary* to be "the science of extracting useful information from large data sets or databases."[xxiv] By partitioning the data into different subsets (i.e., clusters), the data in each subset ideally shares some common trait. Since most ISBN prefixes are uniquely assigned to a single publishing entity for the assignment of full ISBN numbers, the ISBN prefix seemed a good common trait for clustering bibliographic records. Relatively few exceptions occurred in the case of ISBN prefixes assigned to "vanity" presses, and to some publishing communities outside the English-speaking West. However, for the initial research, only English-language cataloging records were examined, and the ISBN prefix was a powerful hook into the overall bibliographic data.

ISBN prefix 019, which belongs to Oxford University Press, was used as a first test group for clustering. In the July 2005 WorldCat snapshot, prefix 019 was the most frequently occurring prefix in WorldCat; it was the prefix for one or more ISBNs within 84,276 records (0.15 percent of WorldCat). The contents of subfield b of MARC field 260 were extracted from these records, and were deemed the publisher name. This process resulted in 91,528 unique strings of text. The publisher names were then normalized according to NACO normalization rules to account for differences in capitalization and punctuation, resulting in 1,550 unique normalized publisher names. The normalized publisher names were clustered using the Levenshtein Distance value.[xxv] The value measures the similarity between two strings by counting the number of deletions, insertions or substitutions of characters needed to transform one string to the other. The publisher strings were then clustered by this distance metric.

Researchers then attempted to automatically resolve the data across the WorldCat database into a set of variant names for each publishing organization. After refinements to the clustering algorithm to better account for noise phrases and punctuation, the ISBN prefixes most frequently appearing in WorldCat as of January 2006 were automatically clustered (see table 2). In the case of the 019 ISBN set, the program yielded an 85 percent agreement. However, an application of the same program to the next four largest groups of ISBN prefixes achieved less than 5 percent success in identifying matches. In other words, this step of the project achieved a workable definition of a number of distinct entities, with their nested inter-relationships, directly from WorldCat data. However, this only was possible with a high level of human intervention.

ISBN prefixes were retained as a data mining technique with different algorithms during the construction of the OCLC PNAF database, as follows. The research team concentrated on a group of high-occurrence publishers. Accordingly, a program was developed to extract sets of ISBN publisher prefixes that represent the highest-occurring ISBN prefixes within the set of database records as sorted by country of publication (The country is defined by its current political boundaries as coded in the MARC fixed field Place of Publication). Researchers constructed a list of the most prominent publishers, seeding the PNAF database with high-incidence publishers from a dozen countries around the world, the top ten research university presses, and any publisher involved in a merger or acquisition during the time of research (under the working assumption that the footprint in the global bibliographic world of any publisher purchasing another would be increasing).[xxvi] A large part of WorldCat was thus clustered according to large subsets of ISBN prefixes.

In the case of each publishing entity identified, an authoritative Preferred Form of the name was first assigned. If the publisher already existed in the NACO National Authority File (http://www.loc.gov/catdir/pcc/naco/nacopara.html) as a corporate name (44 percent of the publishers in the PNAF were included), that authoritative form also was selected for the PNAF Preferred Form. All variant strings mined from the 260 $b bibliographic data were then compared to the Preferred Form; comparisons were made according to a tri-dist fuzzy matching program, and given further manual review afterwards. Tri-dist compares strings based on three-letter sequences called trigrams. To compare two strings, they are typically normalized in some way, for example to eliminate differences in capitalization. Then the two strings to be compared are broken up into overlapping trigrams; for example, using the underscore character to represent a space, the string ‘Al Smith’ generates 8 trigrams: ‘al’, ‘al_‘, ‘l_s’, ‘_sm’, ‘smi’, ‘mit’, ‘ith’, ‘th_’. The trigrams from each string are then compared and a score that estimates the probability of a match is computed based on the proportion of trigrams in common. The fuzzy matches were subjected to human review to assure data quality.

The team then worked outwards from these initial publishing entities, researching all known hierarchical structures related to them, current and past. Relationships were recorded between publishing entities for imprints, acquisitions, and subsidiary divisions, and were collected from a variety of sources. New strings for imprints and related publishing entities were harvested from the 260 $b, and a variety of published business intelligence sources and published company materials were consulted, and each was cited in the PNAF for each instance of source for a data point.

The total number of publishers, imprints, and other publishing entities with records in the PNAF totaled 1,854, with data on their hierarchical relationships and other data (see below, under

Results). To further increase the bibliographic data related to these clusters, the table of all variant strings mapped to each entity was then compared once more to the complete bibliographic database. The verified publisher name strings were compared to all 260 $b contents, to capture records that do not have an ISBN, but that still may be associated with the publisher via the 260 $b field.

This process yielded final data clusters totaling some 16.3 million records, which in turn represent 550 million holdings - slightly more than 33 percent of worldwide library holdings as reflected in WorldCat.[xxvii] This richness is a direct benefit from the decision to begin with high-occurrence entities.

Results

Publisher Name Authority File database

The PNAF database is a relational database capable of management in Microsoft Access. As of this writing, it contains 1,854 records, each representing a single current publishing entity (see list below for definition); and 1,721 records describing the relationships between them, classified by type (Subsidiary division of, Imprint of, Acquired by, Merged with/into, Joint venture with, Re-organized as subsidiary of). The initial entities were identified and researched for inclusion in the database as follows:

- The top twenty-five publishing entities in the United States, as determined by presence of their assigned ISBN prefixes in WorldCat, and the subsidiaries and parents of these entities (see table 3).
- The top twenty publishers in the United Kingdom, and their related entities, determined in the same manner.
- The top ten publishers in Australia, Canada, China, Finland, France, Germany, Italy, Japan, the Netherlands, New Zealand the Russian Federation, Spain, and Taiwan. These

twelve countries together represented more than forty-seven million records in the database, and the initial dataset, mined via ISBN prefixes alone, represented some 3.7 million records.

- The top ten university presses by ISBN prefix in WorldCat.
- Any print publisher involved in a merger or acquisition since November 2003, as reported in the archives of *Publishers' Weekly*.

The team then worked outwards from these initial entities, researching by any means possible – data mining, business intelligence sources (see list of sources in figure 1), Web searching – all known hierarchical structures, current and past; see figure 2 for an example of the complex relationships possible in the 21st-century world of publishers' mergers and acquisitions. All relationships were collected and classified, and the database was built, collecting data on each publisher according to the following fields (see figure 1).

Publisher Name, Preferred Form: The first text field (indexed for searching) contains a single string representing the unique preferred form for one publishing entity. The information in all other fields for the record refers to this entity. The definition between entities (be they holding companies, publishing houses, subsidiary divisions, or distinct imprints) tends to emerge from consideration of the relationships between them, which are classed and recorded in a second data table. The following sources have been consulted for selection of the Preferred Form, in order of precedence:

1) *NACO National Authority File* (NAF), 110 (Corporate Name) field. The NAF 110 file contains approximately 44 percent of all entities indentified in the PNAF database, regardless of nationality. It serves as the first choice for preferred form in order to facilitate inter-operability within Anglo-American cataloging systems, and within cataloging according to the *Anglo-American Cataloging Rules* (AACR2) and *Resource*

Description and Access (RDA).[xxviii] In all cases where the qualifier (Firm) appears in the NAF file, the same string without the qualifier will be added to the Variant Forms field.

2) *Books In Print Online* (W. W. Bowker, accessed via FirstSearch). This is a source that operates closely between the publishing industry and consumers, including libraries; Bowker staff maintain the database rigorously, including telephone follow-up to the legal departments of various larger publishing houses. The Books in Print publisher name database adds a further 37 percent coverage to the publishing entities, and is especially helpful for subsidiary imprints.
3) The *International ISBN Registry* (K. G. Saur, 2004 edition). Between these three most authoritative sources, 93 percent of the publishing entities may be assigned preferred forms.

The final 7 percent or so of the entities, and their associated data to date, were derived from the remaining five sources:

4) *Publishers' Weekly Online*. Though the articles in this journal do not use any controlled language whatsoever, they offer browseable archives and ongoing notifications of mergers and acquisitions.
5) *Hoover's Online*. A business database, offering in many instances valuable information regarding a company's history and some indications of its corporate structure.
6) *Standard and Poor's Corporate Descriptions*. These are only composed for the largest and most important companies, but will include a complete list of subsidiary holdings.
7) The *Directory of Corporate Affiliations* (DIALOG database). This resource includes spotty coverage of publishing entities, but extremely thorough information.
8) Company websites.

The remaining database fields are as follows:

Source of Preferred Form: Citation to one of the above sources.

Former Names (indexed for searching): Earlier forms under which this entity may have published, including earlier corporate names and the full names of some 19th- or early 20th-century publishers whose houses are still in existence. The sources tend to be company histories (in *Hoover's* or on a company website), as well as 510 cross-references if an NAF file exists already. Dates when the name changed, if known, have been included.

Variant Forms (indexed for searching): For each record, current contents include a number of strings that represent variant spellings, common abbreviations, variant known title-page forms, and so on, of the preferred name. The greatest number of strings was mapped into groups for each publishing entity from bibliographic data mining. More than sixty thousand strings have been mapped onto the 1,854 publishing entities in the database.

ISBN Prefixes: This field contains zero or more ISBN prefixes under which a publishing entity releases publications. They are obtained principally from the *ISBN Registry*, but also from *Books in Print* and (rarely) from perusal of an online sale catalog or from other sources. The database as currently built is able to extract ISBN prefixes from all related entities matching specified type(s) and depth(s) of relationship.

HQ City: The principal city in which the entity's headquarters are currently located. Data is derived from any of the above sources. (This should then be the city that most often appears first on title pages, and thus first in the publisher description data, MARC 260 $c.)

HQ Country: The country containing that city.

Other Cities: Other cities in which the entity maintains major publishing (not ordering or distribution) operations.

URL: Unique internet addresses for the entity's commercial website.

The record for Oxford University Press in the PNAF, for instance (see figure 3), contains data entries for each field of interest except relationships (as Oxford has no hierarchical “parents”). Seven important other cities are identified and the number of variant strings associated with this publisher is greater than a thousand.

Unfortunately, the creation of this table of variant strings highlighted the practical limits of automatic parsing of these data. In the case of comparison to the Preferred Form “Oxford University Press,” for instance, the automatic fuzzy matching algorithm – even when correcting for noise words such as articles or frequently appearing words such as “proceedings” – gave a very high match probability to such strings as “Auckland University Press” and even “Harvard University Press,” and a low match probability for strings such as “Published on behalf of the Royal Horticultural Society by Oxford.” The data associated with each publisher at this stage thus still required a large amount of manual review. Research staff then attempted to validate the results, by profiling the subsets of bibliographic data mapped to each publisher.

PNAF Publisher Profiles

Four large clusters within the publisher data were compared, in order to test the robustness of the data partitions being made on the basis of ISBN prefix and publisher description data. Profiles were constructed of the overall publishing footprint of the following four entities:

- *Oxford University Press*: the original cluster of 119,237 bibliographic records with ISBNs became a total data cluster of 210,095 records (0.19 percent of the WorldCat database) when the set of variant strings were mapped back onto the database. That is to say, we added more than ninety thousand records that do not have ISBN data, but are associated with some confidence to Oxford University Press. The manual review step performed on the automatic matches allowed researchers to maintain high confidence in the overall list of variant strings used in this second data capture.
- *Pearson PLC* includes fourteen subsidiaries and acquisitions: an aggregate cluster of 291,433 records (0.27 percent of WorldCat). [xxix] Penguin and its subsidiaries and imprints

were not included in this profile, both to keep the cluster of a comparable size to the other clusters, and to concentrate the profile on the more academic output of Pearson.

- *Springer (Firm)*: 197,263 records (0.18 percent of WorldCat), not including other massive Bertlesmann properties such as Kluwer.
- *Reed Elsevier PLC* (note that this is the form of the name in the *National Authority File* rather than the better-known shorthand "Elsevier"): includes dozens of subsidiaries, with an aggregate cluster of 370,029 records (0.34 percent of WorldCat).

The profiles compared the bibliographic records mapped to these two large publishers and two conglomerates, considering the languages and formats in which they published, as well as the subjects assigned to the published works. Subject analysis was conducted via the three-tiered terminology (i.e., divisions, categories, and subject descriptors) of the OCLC Conspectus to achieve portraits of a publisher's output at different levels of granularity. [xxx] The first feature compared among the four publisher clusters was data on language of publication (as reflected simply in the MARC fixed field). As might be expected, both of the Anglo-centric publishers are dominated by English-language publications (see Appendix). But, from there, of all languages, Latin is second in publication frequency for Oxford, accounting for one in every two hundred works Oxford contributed to WorldCat, while Pearson instead proceeds to Spanish and other modern European languages. Note also Oxford's publications in Middle English and languages spoken in former British colonies. For Springer and Elsevier, on the other hand, both publishers have a strong showing in second (and third) languages beyond English. The data do stem, of course, from a bibliographic database that, although it has surpassed 54 percent non-English cataloging, still tends to represent its Anglo-American cataloging heritage somewhat more heavily (see Appendix). Not surprisingly for a bibliographic database, all four publishers' profiles are dominated in format by printed material, but here as well, Springer has a significantly different profile in electronic content.

At a high level of subject analysis (the thirty-two “Divisions” of the OCLC Subject Conspectus), the profiles continue to demonstrate distinct characters and begin to vary in even more interesting ways. Languages and literature tend to be the most common within global library holdings, followed by history and business.[xxxi] All of these publishers except, notably, Springer, are strong in literature, although Oxford University Press shows the greatest reliance on that field. Oxford’s publication subjects proceed to history, but then music – this indicates the importance of Oxford’s New York office and its emphasis on music publication. Pearson, owner of Cisco and Adobe Presses, on the other hand, skips history in favor of business and then computer science. Springer is heavily dominated by computer science and the harder sciences (with language and literature not even in the top ten), whereas Elsevier’s publications go quickly to law (because they own Butterworths and Martindale-Hubbell) and engineering. Overall, Elsevier has a portfolio that is slightly more balanced among the subjects than any of the other three, as may be seen at each level of subject analysis.

Similarly, at a second more granular specificity of subject analysis (in this case, approximately five hundred subject “Categories”), the four data profiles diverge. Literature and music continue to dominate Oxford’s subject coverage, with history of Britain and the former colonial sphere of South Asia making strong showings. Pearson’s English publications are less in the field of literature than in language arts and education; their strengths in business and computer science also persist to this level of granularity. For Springer and Elsevier, engineering and (in the case of Elsevier, after English literature) law predominate.[xxxii]

At the most granular level, with approximately seven thousand Conspectus “Subjects” available for analysis, the same trends continue. Among the many observations that could be made about the truly focused and granular strengths of each publisher, the 19th century apparently is more important to Oxford University Press than the early 20th, and Shakespeare by himself rises into the top ten subjects (the subject “Bible” is lower at 0.35 percent, not placing it in the top ten). At this level of analysis, Pearson’s primary reliance on English language and education is indicated in the subject areas of the publications. As “Health Professions” is the same subject term at all three levels of analysis, its presence atop the list for Springer might be overstated, but several of the other subjects in Springer’s top publications are as remarkably idiosyncratic as Shakespeare was for Oxford. As noted above, Elsevier (with its immense conglomeration of subsidiaries) maintains the most balanced portfolio of subjects overall: of the approximately 7000 Conspectus subject categories at this third level of granularity, the profile for Elsevier includes publications in 5,630 of them.

From the level of detail and, more importantly, of distinctness in the above profiles, the clusters of bibliographic records that can be associated with each publisher in the PNAF appear robust. Following the success of this result, a prototype graphical interface to the data for all of the publishing entities represented in the PNAF was developed.

WorldCat Publisher Pages Prototype

The often strongly-profiled character of each individual publisher’s group of bibliographic records in WorldCat identified led to the development of a prototype set of web pages based on the PNAF for viewing these data. The prototype also was informed by prior experience with data visualization in the OCLC WorldMap and OCLC Audience Level, prototype services that graphically display global library and book data, and the estimated audience level for library resources, respectively.[xxxiii] Each of the major publishers in the PNAF is given a single webpage, which graphically displays the data profile of their publishing in the global bibliographic universe, as reflected in WorldCat.

Users of the PNAF may navigate either by searching on a publisher’s name (which will key in to both the list of Preferred Forms and the larger list of data-mined strings) or by graphically working through the organizational chart provided for each publisher (see figure 4). Tag clouds allow visual navigation through the profiles of each publisher's author, language, and subject data (figure 5). Graphical interfaces display the Audience Level for the publisher's profile, and the location worldwide of their publications and holdings (figure 6).

Discussion

The automatic methodologies of data mining and clustering enabled researchers to build an experimental database of publisher data. Records were resolved into clusters via ISBN prefixes and via previously-identified publisher name description strings; this process identified the issues associated with the bibliographic data relating to publisher descriptions. Both the automatic parsing of name clusters and the more complex second procedures, which led successfully to the construction of the OCLC Publisher Name Authority File, validated the approach of using ISBN prefix as an initial data element for mining and clustering bibliographic records by publisher. However, in both cases, one lesson learned was that the amount of manual review required hampered research efforts to fully automate the process at a global scale. Researchers had intended to develop a completely automatic process to map publisher name authority information into bibliographic records, and found the task at this point too costly in terms of human intervention. This finding is in line with many earlier projects reported in the literature on the difficulty of fully automating the practice of matching strings to construct authority records.

The planning of the PNAF database as developed in this project included the decision to concentrate on high-incidence publishers, and this decision did yield a very robust dataset to support the PNAF as it stands. The data-mined table of more than sixty thousand variant forms of the 260 $b data allowed more than sixteen million bibliographic records worldwide, representing more than 550 million global holdings, to be mapped to the 1,854 publishers comprising the PNAF database. The large number of strings identified with the cataloging data supports at least one of the theoretical arguments commonly made in favor of authority control: it reduces the amount of data clutter, both the labor-intensive clutter of catalogers entering unregulated strings and the user- and system-unfriendly presence of clutter in the resulting bibliographic records. The data from 260 fields resolved within the PNAF had provided a barrier to access that the resolved form could solve. In addition, the complexity of the hierarchical relationships surrounding many publishers in the current world of mergers and acquisitions makes the organizational chart data an extremely valuable component of the PNAF. Any similar projects or further development of the PNAF can only help librarians better assess their collections by publisher.

The construction of "publisher profiles" verified the methodology by comparing clusters of records assigned to different publishers. The profiles of the subjects, authors, and languages of a publisher's works in the global bibliographic universe as reflected in WorldCat demonstrated in great detail the differences between the clusters of bibliographic records parsed via the PNAF variants. The differences observed tended to fall along predictable lines, given the specific publishers involved. Such profiles of each publisher's footprint in the bibliographic universes as reflected in WorldCat, of course, cannot statistically prove the completeness of these data clusters. However, they offer more detailed and nuanced profiles of the publishers' history than are available anywhere else in the publishing or bibliographic world. At the most granular levels of subject analysis, the profiles offer a detailed picture of a publisher's character and, pointedly, each of the four publishers' characters were observed to be quite distinct.

These differences also tend to validate the unique intelligence present within the PNAF data: librarians, publishers, and users can view a portrait of the publisher's output in terms of the authors most associated with the publisher, the languages published, and most importantly, the subjects in which a publisher offers the most expert concentrations. The profiles can be valuable, as expected in the research goals, to fill the need for name authority control within ERM systems to inform collection analysis and development and approval plans, and to publishers as they consider their competitive position in the library marketplace as well as to aid users and public service librarians in discovering publisher outputs. As stated by Blake and Samples, "…OCLC's publisher name authority server nonetheless demonstrations [sic] a need for organization name authorities and may provide context for librarians whose methods and research have already prompted similar projects."[xxxiv]

The value of the publisher profiles led to the construction of the WorldCat Publisher Pages. All of the publishers in the experimental PNAF database are represented by a single Web page, containing their data from PNAF (including Preferred Form of the name, cities of operation, and most importantly, the hierarchical organization chart) and their publication profile in global WorldCat, authors, languages, and subjects.

For future work, researchers have been considering recommendations on ways to code the authorized form of publisher names directly into MARC records, if a completely automatic process for resolving the names could be developed. The most obvious place would be MARC field 710 Corporate Name Added Entry, with the publisher name perhaps occupying a new $6, with NACO in $2 where the Preferred Form of the Name also may be linked to the NACO Authority file (44 percent of the current PNAF). The place of publication, if it could be similarly standardized in the future, could occupy a 752 Added Entry Hierarchical Place Name, with $2 for the FAST terminology currently embedded in PNAF place names.[xxxv] The database currently uses its own unique identifiers, but researchers have been in discussion with those developing the International Standard Name Identifier (ISNI) system regarding incorporation of the PNAF publisher names. However, the reliance on human intervention to update and maintain the database is a detriment to inclusion in other systems and services at this time.

Conclusion

The present research on publisher names both confirmed them as a difficult issue in Anglo-American cataloging, and set a potential example towards providing authority control over them.

The project set out to construct a database containing authoritative strings for publisher names as well as a variety of data relating to their publication output, and accomplished this goal. Each automatic methodology worked to generate clusters of items based on an assigned publisher, first via ISBN prefix and then via further matches of 260 $b data, leading to a robust database of high-incidence publishers. Though the process could not be fully automated on a global scale, some 1,854 high-impact publishing entities were profiled by their publishing output, with detailed differences emerging between the profiles. The profiles as a research output are freely available on the Web via the WorldCat Publisher Pages.

The data captured for each publisher provide a model service for advanced collection analysis and provide additional value for user access to library resources. Tens of thousands of variant strings were resolved to the small number of publishers in the database, potentially reducing cataloging time by providing automatic suggestion of Preferred Forms for publisher names to catalogers. Further applications of this authority control procedure in the Publisher Description could code the Preferred Form of the publisher name directly into the MARC records even if a fuller, more informative string were entered in the publisher description area. Such an application of authority control, even for the limited number of (high-impact) publishers in the PNAF, would offer benefits to both publishers and collection development librarians, by increasing the power of collection analysis tools to parse a collection by publishing agency, and to users and academic public service librarians by allowing better access to searches by publisher name.

Appendix

PNAF Publisher Profiles

Oxford University Press:	%	Pearson PLC:	%	Springer (Firm):	%	Reed Elsevier PLC:	%
Language Data							
English	96.74	English	95.27	English	61.25	English	83.64
Latin	0.51	Spanish	1.43	German	37.10	French	9.34
German	0.39	German	1.33	French	1.02	Dutch	2.32
Chinese	0.39	French	0.60	Italian	0.29	Spanish	0.95
French	0.37	Dutch	0.55	Polish	0.13	Italian	0.60
Spanish	0.28	Latin	0.26	Czech	0.04	Latin	0.27
Afrikaans	0.14	Malay	0.06	Spanish	0.04	Afrikaans	0.16
Middle English	0.13	Ancient Greek	0.05	Hungarian	0.03	Ancient Greek	0.12
Malay	0.09	Portuguese	0.05	Dutch	0.02	Portuguese	0.09
Swahili	0.09	Italian	0.04	Danish	0.02	Polish	0.06
Format Data							
Print	89.57	Print	92.98	Print	81.69	Print	92.31
Computer	8.23	Microform	2.82	Computer	17.51	Computer	5.46
Microform	1.39	Computer	2.15	Microform	0.71	Microform	1.85
Audio	0.50	Video	0.70	Video	0.05	Video	0.14
Video	0.16	Audio	0.67				
Subject Division Data							
Language & literature	27.12	Language & literature	18.67	Computer science	16.83	Language & literature	14.18
History	11.92	Business & economics	13.30	Engineering	15.12	Law	11.78
Music	9.78	Computer science	9.42	Math	12.96	Engineering	11.73
Philosophy & religion	9.55	Engineering	8.04	Medicine	9.93	Business & economics	6.82
Business & economics	6.15	History	7.59	Physical sciences	9.83	Medicine	6.50
Medicine	4.36	Math	6.04	Biology	5.22	Physical sciences	5.01
Law	3.85	Education	5.64	Business &	5.13	History	4.57

Sociology	3.75	Sociology	4.18	economics Health professions	4.48	Biology	4.32
Political science	3.58	Philosophy & religion	3.81	Chemistry	3.14	Health professions	3.70
Biology	2.60	Physical sciences	2.75	Geography	2.58	Chemistry	3.51
Subject Category Data							
English literature	10.66	English language	7.74	Computer science	5.23	English literature	5.84
English language	5.86	Business admin.	4.62	General math	4.48	Health professions	3.40
Microform	1.39	Computer	2.15	Microform	0.71	Microform	1.85
Instrument. Music	3.48	English literature	3.63	Health professions	4.03	English language	2.79
Vocal music	3.09	Economics	2.94	Electrical engineering	3.73	U.S. Federal law	2.32
Literature on music	2.26	Computer program.	2.39	General engineering	3.25	General engineering	2.26
History – Britain	1.82	Electrical engineering	2.24	Mathematic analysis	3.06	Electrical engineering	2.10
Economic history	1.38	Early child-hood ed.	2.05	Computer software	2.37	General law	1.70
American literature	1.35	Computer software	1.88	Computer program.	2.34	Industrial economics	1.65
History – S. Asia	1.30	U.S. Federal law	1.80	Probability/ statistics	2.20	Business admin.	1.53
General history	1.29	Computer science	1.54	Mechanical engineering	2.17	U.S. State law	1.46
Third-level Subject Data							
English – modern	5.57	English – modern	7.68	Health professions	3.56	English – modern	2.68
English lit. – prose	2.51	Manage-ment	2.53	Math collections	2.76	English lit. – prose	2.06
English lit. – 19th cent.	2.23	Program-ming	1.74	Computer science	1.84	Health professions	1.92
Juvenile literature	1.06	Arithmetic	1.09	Program-ming	1.46	U.S. State law	1.37
English lit. – poetry	1.03	Economic theory	1.06	Access/ security	1.10	Industrial managemt.	1.22
English lit. – collected	0.80	Marketing	1.06	Artificial intelligence	1.03	Legal periodicals	1.16
Biographies	0.76	General algebra	1.04	Mathematic statistics	1.03	English lit. – 1900-60	1.15
English lit. – 1900-60	0.74	Accounting	0.97	Analytical physics	1.02	Engineering materials	0.86
Shakes-peare	0.68	Juvenile literature	0.93	Industrial managemt.	0.99	English fiction	0.83
Sacred choruses	0.66	English lit. – 19th cent.	0.89	Engineering materials	0.90	Nuclear physics	0.68

Tables

Table 1: Changes to Cataloging Rules for Multiple Places and Publishers

Prominence of Place/ Publisher	*ALA Cataloging Rules (1941)*	*Anglo-American Cataloguing Rules (1967)*	*Anglo-American Cataloguing Rules, revised (2002)*
Neither is prominent	-First listed first -Indicate omission	-First listed only -Omit others	-First listed only -Omit others
First is prominent	-First listed first -Indicate omission	-First listed only -Omit others	-First listed only -Omit others
First is not prominent	-Prominent listed first -First listed second	-Prominent listed only -Omit others	-First listed first -Prominent listed second

American Library Organization, Catalog Code Revision Committee, *A.L.A. Cataloging Rules for Author and Title Entries* (Chicago: ALA, 1941).

American Library Organization, *Anglo-American Cataloguing Rules* (Chicago: ALA, 1967).

Joint Steering Committee for Revision of AACR, *Anglo-American Cataloguing Rules 2002 Revision* (Chicago: ALA, 2003).

Table 2: Automatic Parsing Summary

Prefix	***WorldCat Records***	***Unique 260 $b Strings***	***Program-assigned strings***	***Strings requiring review***	***% strings requiring review***
0-19	101,347	2,089	1,788	301	14.41
0-315	100,619	500	1	499	99.80
0-612	97,284	219	0	219	100.00
0-665	88,301	14,260	83	14,177	99.42
0-13	68,125	2,148	75	2,073	96.51

These data are from January 2006, and include all records associated with the respective ISBN prefixes published worldwide.

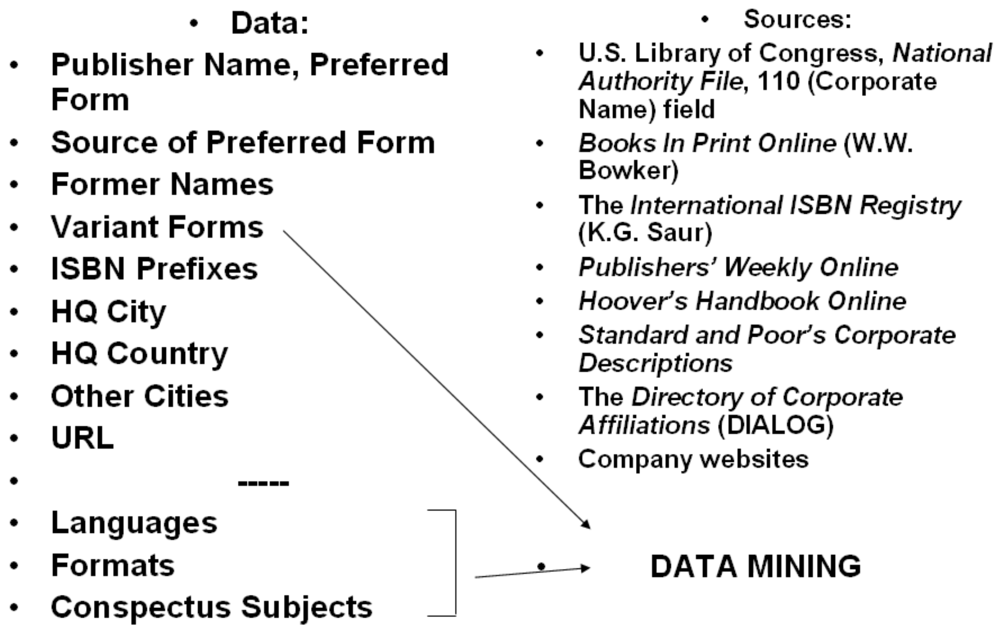
Table 3: Example of high-incidence ISBN prefixes for Publications from the United States in WorldCat

ISBN Prefix	*WorldCat Records*	*Publishing Entity, PNAF Preferred Form*
0-13	50,298	Prentice-Hall, Inc.
0-07	44,545	McGraw Hill, Inc.
0-06	44,362	HarperCollins (Firm)
0-16	40,451	United States G.P.O.
0-471	37,710	John Wiley & Sons
0-312	33,318	St. Martin's Press
0-671	31,765	Simon & Schuster, Inc.
0-02	27,602	MacMillan Publishers
0-15	18,420	Harcourt Brace & Company
0-394	18,043	Random House (Firm)
0-590	17,290	Scholastic Inc.
0-385	16,768	Doubleday and Company, Inc.
0-395	16,699	Houghton Mifflin Company
0-19	15,724	Oxford University Press
0-03	15,417	Holt, Rinehart, and Winston

These data are from March 2006, in the initial stages of the OCLC Publisher Name Authority File methodology, and refer to works with the United States as the country of publication.

Figures

Figure 1: Structure and sources for the OCLC Publisher Name Authority File



- Data:
- Publisher Name, Preferred Form
- Source of Preferred Form
- Former Names
- Variant Forms
- ISBN Prefixes
- HQ City
- HQ Country
- Other Cities
- URL
- -----
- Languages
- Formats
- Conspectus Subjects

- Sources:
- U.S. Library of Congress, *National Authority File*, 110 (Corporate Name) field
- *Books In Print Online* (W.W. Bowker)
- The *International ISBN Registry* (K.G. Saur)
- *Publishers' Weekly Online*
- *Hoover's Handbook Online*
- *Standard and Poor's Corporate Descriptions*
- The *Directory of Corporate Affiliations* (DIALOG)
- Company websites

- DATA MINING

Figure 2: The OCLC Publisher Name Authority File complex of subsidiaries for Pearson PLC

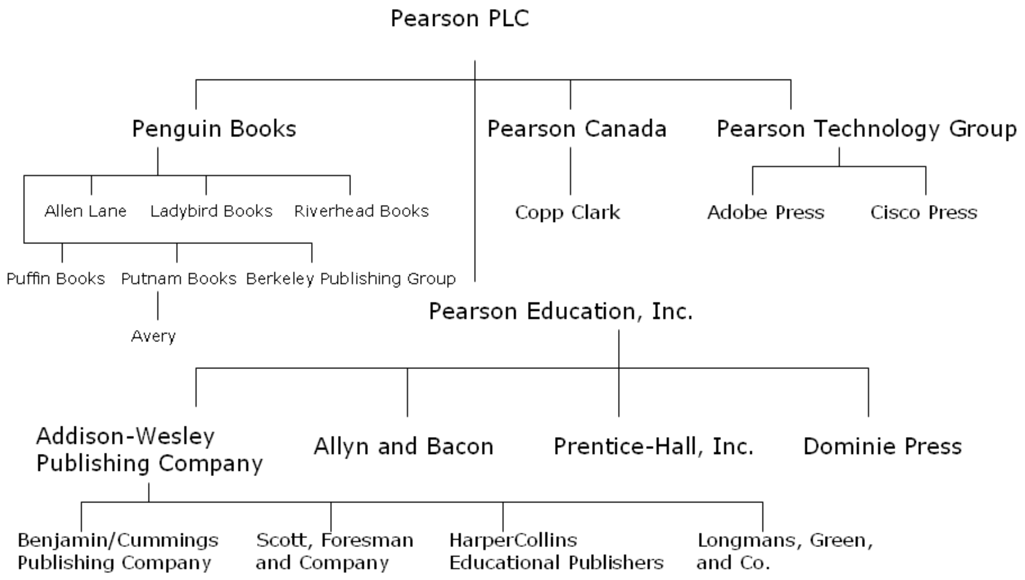
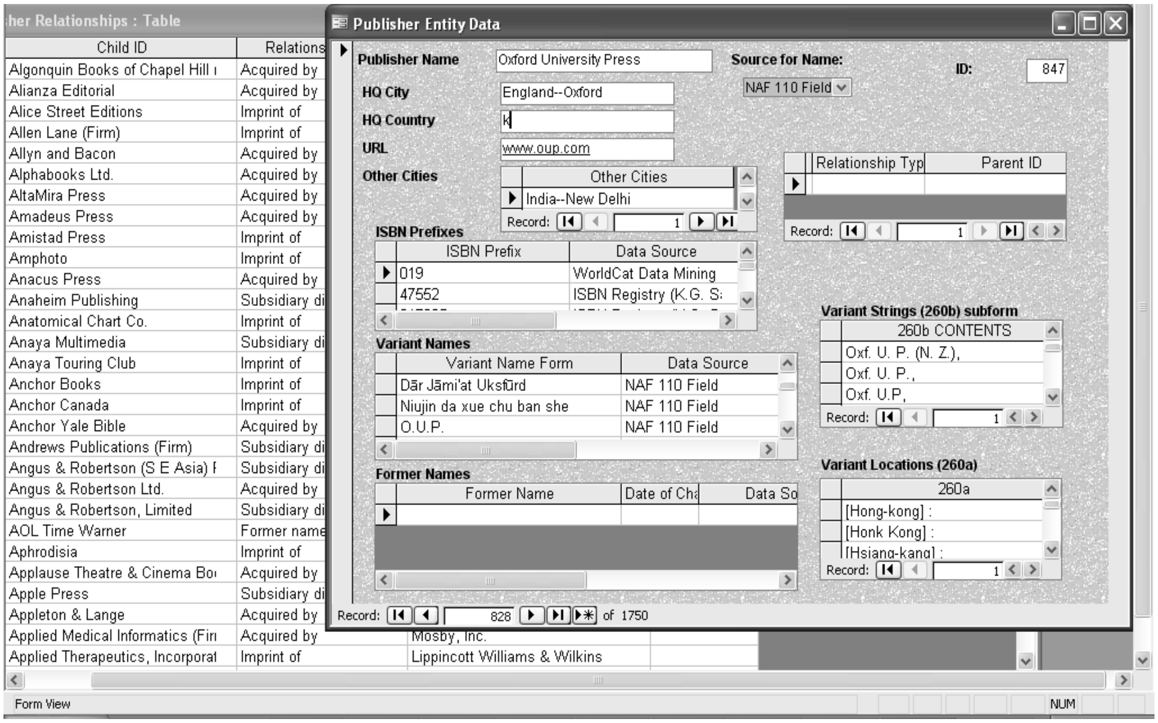


Figure 3: Oxford University Press record in the OCLC Publisher Name Authority File



her Relationships : Table

Child ID	Relations	
Algonquin Books of Chapel Hill ı	Acquired by	
Alianza Editorial	Acquired by	
Alice Street Editions	Imprint of	
Allen Lane (Firm)	Imprint of	
Allyn and Bacon	Acquired by	
Alphabooks Ltd.	Acquired by	
AltaMira Press	Acquired by	
Amadeus Press	Acquired by	
Amistad Press	Imprint of	
Amphoto	Imprint of	
Anacus Press	Acquired by	
Anaheim Publishing	Subsidiary di	
Anatomical Chart Co.	Imprint of	
Anaya Multimedia	Subsidiary di	
Anaya Touring Club	Imprint of	
Anchor Books	Imprint of	
Anchor Canada	Imprint of	
Anchor Yale Bible	Acquired by	
Andrews Publications (Firm)	Subsidiary di	
Angus & Robertson (S E Asia) I	Subsidiary di	
Angus & Robertson Ltd.	Acquired by	
Angus & Robertson, Limited	Subsidiary di	
AOL Time Warner	Former name	
Aphrodisia	Imprint of	
Applause Theatre & Cinema Bo	Acquired by	
Apple Press	Subsidiary di	
Appleton & Lange	Acquired by	
Applied Medical Informatics (Fir	Acquired by	Mosby, Inc.
Applied Therapeutics, Incorporat	Imprint of	Lippincott Williams & Wilkins

Form View

NUM

Publisher Entity Data

Publisher Name: Oxford University Press

Source for Name: NAF 110 Field

ID: 847

HQ City: England--Oxford

HQ Country: k

URL: www.oup.com

Other Cities

Other Cities
India--New Delhi

Record: 1

Relationship Typ	Parent ID

Record: 1

ISBN Prefixes

ISBN Prefix	Data Source
019	WorldCat Data Mining
47552	ISBN Registry (K.G. S:

Variant Names

Variant Name Form	Data Source
Dār Jāmi'at Uksfūrd	NAF 110 Field
Niujin da xue chu ban she	NAF 110 Field
O.U.P.	NAF 110 Field

Former Names

Former Name	Date of Cha	Data So

Variant Strings (260b) subform

260b CONTENTS
Oxf. U. P. (N. Z.),
Oxf. U. P.,
Oxf. U.P,

Record: 1

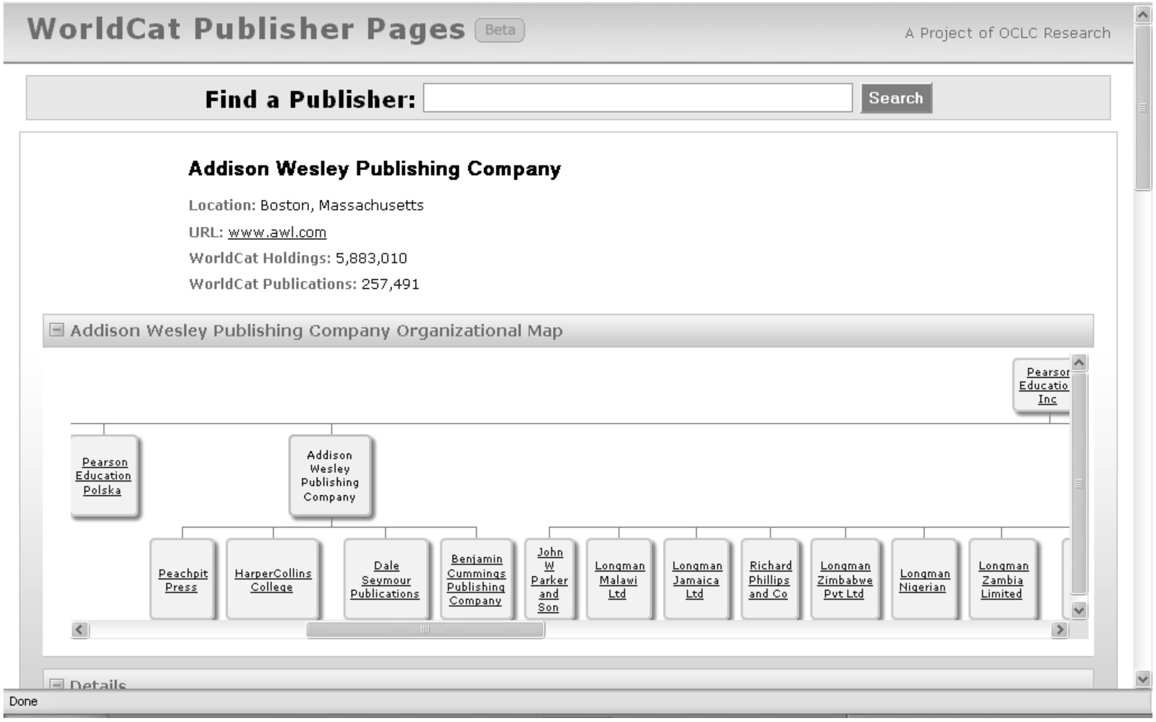
Variant Locations (260a)

260a
[Hong-kong] :
[Honk Kong] :
[Hsiang-kang] :

Record: 1

Record: 828 of 1750

Figure 4: WorldCat Publisher Page, with Organizational Chart



Addison Wesley Publishing Company

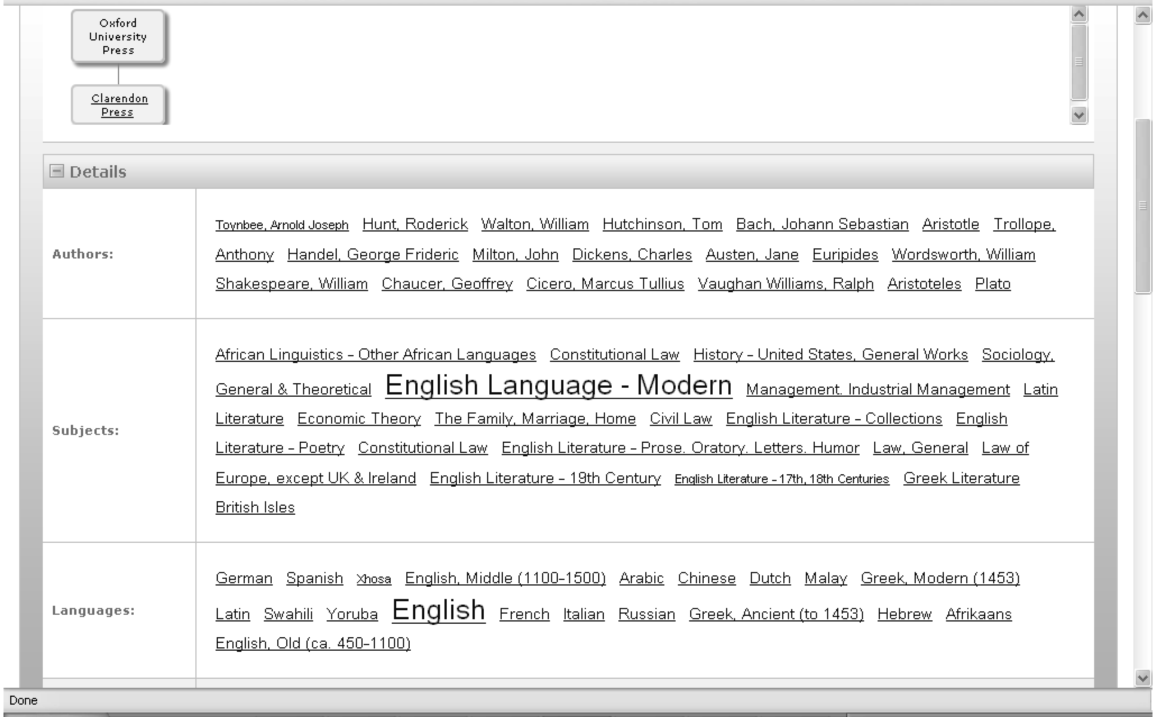
Location: Boston, Massachusetts

URL: www.awl.com

WorldCat Holdings: 5,883,010

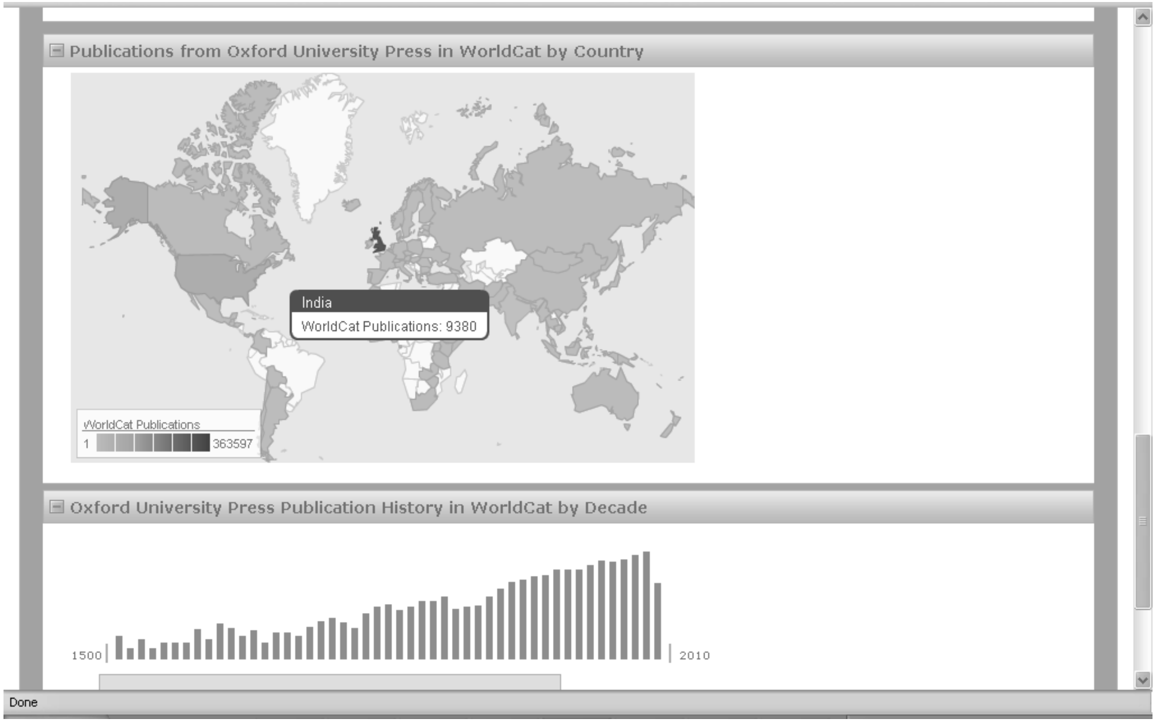
WorldCat Publications: 257,491

Figure 5: WorldCat Publisher Page, with Profile Data



Authors:	Toynbee, Arnold Joseph Hunt, Roderick Walton, William Hutchinson, Tom Bach, Johann Sebastian Aristotle Trollope, Anthony Handel, George Frideric Milton, John Dickens, Charles Austen, Jane Euripides Wordsworth, William Shakespeare, William Chaucer, Geoffrey Cicero, Marcus Tullius Vaughan Williams, Ralph Aristoteles Plato
Subjects:	African Linguistics - Other African Languages Constitutional Law History - United States, General Works Sociology, General & Theoretical English Language - Modern Management. Industrial Management Latin Literature Economic Theory The Family, Marriage, Home Civil Law English Literature - Collections English Literature - Poetry Constitutional Law English Literature - Prose. Oratory. Letters. Humor Law, General Law of Europe, except UK & Ireland English Literature - 19th Century English Literature - 17th, 18th Centuries Greek Literature British Isles
Languages:	German Spanish Xhosa English, Middle (1100-1500) Arabic Chinese Dutch Malay Greek, Modern (1453) Latin Swahili Yoruba English French Italian Russian Greek, Ancient (to 1453) Hebrew Afrikaans English, Old (ca. 450-1100)

Figure 6: WorldCat Publisher Page, with Publication Data



Publications from Oxford University Press in WorldCat by Country

Oxford University Press Publication History in WorldCat by Decade

References and Notes

[i] Nirmala S. Bangalore and Chandra Prabha, “Authority Work in Copy (Derived) Cataloging: A Case Study,” *Technical Services Quarterly* 15, no. 4 (1998): 54.

[ii] Barbara B. Tillett, “Authority Control: State of the Art and New Perspectives,” *Cataloging & Classification Quarterly* 38, no. 3/4 (2004): 23-41 at 24.

[iii] Tillett, “Authority Control,” 24; for an extensive review of the literature on authority control from 1900, see Larry Auld, “Authority Control: An Eighty-Year Review,” *Library Resources & Technical Services* 26, no. 4 (1980): 319-30. For a good, recent overview of resources, see Robert E. Wolverton, Jr., “Becoming an Authority on Authority Control: An Annotated Bibliography of Resources,” *Library Resources & Technical Services* 50, no. 1 (January 2006): 31-41.

[iv] Arlene G. Taylor, *Wynar’s Introduction to Cataloging and Classification*, 9th ed. rev. (Westport, Conn.: Libraries Unlimited, 2004), 491.

[v] The group of studies appear in *Cataloging & Classification Quarterly* 39/1-2 (2004).

[vi] M. M. M. Synman and M. Jansen van Rensburg, “Reengineering Name Authority Control,” *Electronic Library* 17, no. 5 (1999): 313-22.

[vii] Marielle Veve, “Supporting Name Authority Control in XML Metadata: A Practical Approach at the University of Tennessee,” *Library Resources & Technical Services* 53, no. 1 (Jan. 2009): 41-52.

[viii] Mark Patton et al., “Toward a Metadata Generation Framework,” *D-Lib* 10 no. 11 (Nov. 2004), www.dlib.org/dlib/november04/choudhury/11choudhury.html (accessed Feb. 1, 2011).

[ix] Marko Rodriguez, Johan Bollen, and Herbert van de Sompel, “Automatic Metadata Generation Using Associative Networks,” *ACM Transactions on Information Systems* 27, no. 2 (April 2009): 1-20.

[x] Kristen Blake and Jacquie Samples, “Creating Organization Name Authority within an Electronic Resources Management System,” *Library Resources & Technical Services* 53, no. 2 (April 2009): 94-107.

[xi] See John D. Bynum, Jr., “NACO: A Cooperative Model for Building and Maintaining a Shared Name Authority Database,” *Cataloging & Classification Quarterly* 38, no.3/4 (2004): 237-49.

[xii] James C. French, Allison L. Powell, and Eric Schulman, “Using Clustering Strategies for Creating Authority Files,” *Journal of the American Society for Information Science* 51, no. 8 (2000): 774-86.

[xiii] Ibid, 776.

[xiv] Lynn Silipigni Connaway and Akeisha Heard, “Publisher Name Authority Project: An Attempt to Enhance Data Mining for Collection Analysis and Comparison,” paper presented at the XXV Annual Charleston Conference, Charleston, SC, 4 November 2005.

[xv] *Anglo-American Cataloguing Rules*, 2nd ed.., 2002 rev. (Ottawa: Candian Library Assn.; Chicago: ALA, 2002rule 1.4D2 (.

[xvi] Qiang Jin, “Comparing and Evaluating Corporate Names in the National Authority File (LCNAF) on OCLC and on the Web,” *Cataloging & Classification Quarterly* 36, no. 2 (2003): 21-31.

[xvii] See E. Kasinec and R. H. Davis, “Materials for the Study of Russian/Soviet Art and Architecture: Problems of Selection, Acquisition, and Collection Development for Research

Libraries in Historical Perspective," *Art Documentation* 10, no. 1 (Spring 1991): 19-22, for one example.

[xviii] Lynne Branche Brown, "Standards for Acquisitions Data: Report of the ALCTS Automated Acquisitions Discussion Group Meeting, American Library Association Annual Conference, Toronto 2003," *Technical Services Quarterly* 21, no. 3 (2004): 79-81.

[xix] See OCLC Research, www.oclc.org/research/activities/past/orprojects/mining/default.htm (accessed Nov. 24, 2010), for a partial bibliography of studies.

[xx] Brian Lavoie, Lynn Silipigni Connaway, and Edward T. O'Neill, "Mapping WorldCat's Digital Landscape," *Library Resources & Technical Services* 51, no. 2 (April 2007): 106-15.

[xxi] Lynn Silipigni Connaway and Timothy J. Dickey, "Beyond Data Mining: Delivering the Next Generation of Service from Library Data," [part of the panel presentation "Transforming Data into Services: Delivering the Next Generation of User-Oriented Collections and Services"], *Proceedings of the American Society for Information Science & Technology*, 45, no. 1 (2008): 1059-63 at 1062. This is incomplete

[xxii] OCLC, Worldcat Facts and Statistics, www.oclc.org/us/en/worldcat/statistics/default.htm (accessed Feb. 15, 2011); percentage of non-English records (cited as a 2010 statistic) from Jay Jordan, "OCLC Update Breakfast at ALA Midwinter 2011," www.oclc.org/news/events/presentations/default.htm (accessed Feb. 15, 2011).

[xxiii] Connaway and Heard, "Publisher Name Authority." Please move this sentence to the text of the paper.

[xxiv] The Free Dictionary, "Clustering," http://encyclopedia.thefreedictionary.com/clustering (accessed Nov. 24, 2010).

[xxv] Mike Gilleland, "Levenshtein Distance," www.merriampark.com/ld.htm (accessed Nov. 24, 2010).

[xxvi] Mergers and acquisitions were taken as reported in *Publisher's Weekly*, between January 2001 and October 2009; see Lynn Silipigni Connaway and Timothy J. Dickey, "Beyond Data Mining: Delivering the Next Generation of Library Services." Paper presented at the Annual Meeting of the American Society for Information Science & Technology, Columbus, Ohio, Oct. 28, www.oclc.org/research/presentations/connaway/asist2008.ppt (accessed Nov. 24, 2010).

[xxvii] Jeremy Browning, e-mail to Timothy J. Dickey, Aug. 11, 2010.

[xxviii] *Anglo-American Cataloguing Rules*; Joint Steering Committee for Development of RDA, "RDA: Resource Description and Access," http://www.rda-jsc.org/rda.html (accessed Feb. 15, 2011). See also Chris Oliver, *Introducing RDA: A Guide to the Basics* (Chicago: ALA, 2010).

[xxix] The profiles were constructed from data in September of 2008, and were quickly made slightly obsolete in the fast-paced world of publisher's merger and acquisition activity. The data for the group of publishers grouped together under Pearson do not include the publications of Heinemann-Raintree, which was sold to Pearson in 2009; *Publisher's Weekly*, 255/39 (September 29, 2008): 4-6. The sub-division Heinemann-Raintree Reference Library was soon after re-sold to Capstone Publishing.

[xxx] OCLC, Introduction to the WorldCat Collection Analysis Services, "1. 2 The OCLC Conspectus," www.oclc.org/us/en/support/documentation/collectionanalysis/using/introduction/introduction.htm (accessed Nov. 24, 2010).

[xxxi] Edward T. O'Neill and Julia A. Gammon, "Building Collections Cooperatively: Analysis of Collection Use in the OhioLINK Library Consortium," in *Pushing the Edge: Explore, Engage, Extend: Proceedings of the Fourteenth National Conference of the Association of College and research Libraries March 12-15, 2009, Seattle, Washington*, ed. Dawn M. Mueller (Chicago: ACRL, 2009): 36-45, www.aasl.org/ala/mgrps/divs/acrl/events/national/seattle/papers/36.pdf, (accessed Nov. 24, 2010).

[xxxii] "Health Professions" is the same term on all three levels of Conspectus subject analysis, so it will filter upwards at the second and third levels of specificity.

[xxxiii] On the OCLC Audience Level, see Edward T. O'Neill, Lynn Silipigni Connaway, and Timothy J. Dickey, "Estimating the Audience Level for Library Resources," *Journal of the American Society for Information Science & Technology* 59. No. 11 (Nov. 2008): 2042-50. See OCLC, "WorldMap," www.oclc.org/research/activities/worldmap/default.htm (accessed Nov. 24, 2010) and OCLC, "Audience Level," www.oclc.org/research/activities/audience/default.htm (accessed Nov. 24, 2010) for more information; the services are available at http://www.oclc.org/globallibrarystats/default.htm and http://audiencelevel.oclc.org/AudienceLevel/al, respectively (all accessed Nov. 24, 2010).

[xxxiv] Blake and Samples, "Creating Organization Name Authority," 97-98.

[xxxv] OCLC, "FAST (Faceted Application of Subject Terminology," www.oclc.org/research/activities/fast/default.htm (accessed Nov. 24, 2010.)