Explicit and Implicit Searching in the Perseus Digital Library

Anne Mahoney
Perseus Project, Tufts University
amahoney@perseus.tufts.edu

The Perseus Digital Library (http://www.perseus.tufts.edu) has three different kinds of search tools, each of which presents its results in a different way. Two of them, the word-search tool and the Perseus Lookup Tool, allow users to request searches of the texts and the databases of the digital library. The third search tool works transparently on the user's behalf, and is tightly integrated into the text display engine. In this paper, I will describe these three search tools. I will focus on the third, the automatic search mechanism, as a case study in an integrated "information push" strategy.

The original focus of the Perseus project[1] was on Ancient Greek texts, art, archaeology, and history. The project has been expanded to include Latin and Roman materials as well as Early Modern English, and we are considering additional projects in European and American history and literature. Perseus includes all the classical texts commonly read in American schools and colleges, all in the original language and most also in English translation, as well as less common authors such as Apollodorus and Strabo. The automatic search mechanism described here makes the more obscure authors accessible to readers by offering passages relevant to what is currently being read.

1. Looking up words

The most straightforward of the Perseus search facilities is the suite of word-search tools, similar to those found in most first-generation text bases. These tools allow users to search for words or phrases, in Latin, Greek, or English, anywhere in the texts of the library or in dictionary entries. The Latin and Greek word search tools find words in the Latin or Greek texts, allowing the user to specify one or more words to be included in or excluded from the search. These words may be inflected forms or standard citation forms. Normally, we search for any form of the same word, but users may also specify that they want an exact match for the particular form entered. Users also indicate which texts they want to search, by genre or by author. The results appear as a list of sentences extracted from the texts, in which the target words are highlighted in red. The title, author, and location are given for each sentence, with a link to the passage in its full context. The results can appear either ordered by the name of the author of the work or ranked by goodness of match.

Goodness of match means not only the aggregate precision – how many of the user's specified terms appear in the result sentence –, but also how certain we are of whether the matching forms really match. Because of the complexity of Latin and Greek morphology, it is not always possible to identify unambiguously the word to which a form belongs. The search results page marks ambiguous forms with an asterisk, but does not attempt to indicate the range of possible dictionary entries from which the given form may have come. An example in English may clarify the principle: *forms* might be a noun, in a phrase like

[1] The Perseus project has been under development since 1987, and Perseus materials have been published in several formats. Since 1996, the project has been available on the World Wide Web; currently, the main site and its European mirrors receive some two million hits a week. More detailed descriptions of the Perseus Project can be found in Crane, "Perseus Project," and Smith et al.

"forms of government," or a verb, in a phrase like "the baker forms the loaf." In Latin, similarly, *formas* might be the accusative plural of the noun *forma* 'a form' or the second person singular present indicative active of the verb *formare* 'to form.' In many cases, as in this example, the possible words are similar in meaning. In other cases, they are not: words as dissimilar semantically as *auris* 'ear', *aurum* 'gold', and *aura* 'breeze' can have identical forms. As these are all nouns, moreover, they cannot be told apart on syntactic grounds. We therefore leave the disambiguation to the human reader.[2]

The word-search tool kit also includes a dictionary lookup tool. In addition to ordinary lookups by headword (Latin words in the Latin dictionary, Greek in the Greek), we allow users to look up English words in the definitions. That is, a user can search for a Latin word meaning 'infantry' – more precisely, a word whose dictionary entry contains 'infantry'. Searches may be restricted to words used by particular authors.

Results of the dictionary searches are presented in table form, alphabetical by word. The first column contains the resulting dictionary headwords, each linked to its dictionary entry. Subsequent pairs of columns give frequency information: the total number of occurrences and the frequency of the word in the digital library, calculated as instances per 10,000 words. The first pair gives these statistics on the assumption that all ambiguous forms are counted as belonging to this word ("maximum instances"); the next pair assumes none of the ambiguous forms belongs to this word ("minimum instances"). The numbers are links to the word-search tool described above, and to a word-frequency tool, which shows the number of times the target word appears in various sub-corpora of the digital library, including in particular the works of individual authors.

The word-search tools allow users to explore how particular words are used by particular Greek or Latin authors, what words are used together, and what words belong to overlapping semantic fields.

2. The Lookup Tool

The Perseus Lookup Tool is the second kind of search. It allows users to search for keywords anywhere in the digital library: in texts, in the image database, in the atlas. Fuzzy matches and alternate name tables permit searches even for Greek names that may be transliterated in more than one way. This general facility allows users to request everything in the library about a particular divinity, city, historical figure, author, or the like.

Results from the Lookup Tool appear as a collapsible list, grouped by type of information: all the pictures of vases are listed together, as are all the entries in the Perseus Encyclopedia, and so on. Each group in the list indicates how many items it contains. If there is only one item, its description is shown and linked to the item itself. If there are several matching items in a particular category, the group is shown in collapsed form. For vases, sculptures, coins, buildings, and general images, a button next to the group name offers a link to the Image Browser for thumbnail images.

The Perseus Lookup Tool can provide an overwhelming amount of information for the most important keywords. For Athens, for example, we show:

- 2 Atlas sites
- 53 buildings
- 4 articles in the Caskey-Beasley catalog of vases in the Museum of Fine Arts, Boston

[2] Further information on the morphological analysis tools within Perseus can be found in Crane, "Generating and Parsing Ancient Greek."

- 28 coins
- 9 sections in the overview of Greek history by Thomas Martin
- 1020 images of places in and around Athens
- 71 entries in the catalog of document reliefs by Carol Lawton
- 1 entry in the Princeton Encyclopedia of Classical Sites
- 896 sculptures
- 3 sites
- 76 bibliographic references
- 16 entries in the overview of Greek sculpture by Andrew Stewart
- 1 text
- 102 vases
- 1 work

for a total of 2283 references. Grouping them by type allows users to select textual or pictorial information; it also keeps the on-screen display to a manageable size. There is no attempt to rank the relevance of the results: we simply provide every object in the digital library whose description contains the search term. Users must then choose among geographic, artistic, historical, or literary objects, and must choose primary sources (texts, images) or secondary (encyclopedias and overview texts). Those users who already have a clear idea of the importance of, say, Athens can quickly pick out the information relevant to their present projects. Relative novices, however, receive little guidance from this tool about what sources are most likely to be useful.

3. Automatic, integrated searching

The most interesting searches within Perseus, however, are the ones the user does not have to request. Whenever Perseus displays a text, relevant information from elsewhere in the digital library is automatically extracted and made available, in the form of hyperlinks within or next to the text being read. This automatic search mechanism includes both the features of the Lookup Tool and connections between the text being read and any other in the digital library that quotes it. That is, citations within the Perseus Digital Library form *bi-directional* links: texts "hear" other texts quoting them. Readers recognize that additional information is available, and can choose whether or not to follow the links, just as they may choose whether or not to read footnotes in a printed text. Readers can also configure how the intertextual links are displayed and what classes of links are presented.

Within any text in English, keywords known to the Lookup Tool are automatically linked to that search facility; this is not currently done for Greek or Latin texts because the Lookup Tool keyword list is in English, and because Greek and Latin words are customarily linked to the morphological analysis tool. Lookup links furnish identifications of mythological and historical characters and of geographic names. They act like the glossary at the back of a printed translation, but with the difference that readers always know whether a particular term is glossed or not.

When a document mentions places, Perseus offers one or more links to the Perseus Atlas.[3] Readers can plot all the sites in the entire text or in the page being read. For works divided into "books," readers can also plot all the sites mentioned in the current book. If a user follows an Atlas search link, the Atlas opens in a separate browser window, to facilitate comparing the map and the text.

Intertextual links apply to texts in any language. Whenever one text in the digital library cites another, links appear both at the point of the citation and at the cited point. For

[3] The atlas is described in Chavez, "Geospatial Data."

example, suppose a reference grammar cites a text as an example of a particular syntactic or morphological feature, as Smyth's *Greek Grammar for Colleges* frequently cites Xenophon's *Anabasis*. In the grammar, this reference appears as a link to the text of Xenophon; this is straightforward, and typical of on-line editions generally. The citation in the grammar manifestly points to Xenophon, in any edition of Smyth's grammar, printed or on line.

A printed text of Xenophon, however, cannot "know" that it is cited by Smyth's grammar – but an on-line text can. A reader of Xenophon in the Perseus Digital Library will see when a given passage is used as an example in the grammar. The first page of the *Anabasis*, for example, is quoted or mentioned some 50 times in Smyth's grammar. Each quotation is linked back to the grammar with a superscript cross within the text; the discussions that mention, but do not quote, the text are listed at the bottom of the page of Xenophon, also with links to the grammar. Someone reading this text can see at once which passages are discussed elsewhere.

While the technique of co-citation analysis is familiar, the presentation of these results within the Perseus Digital Library is unusual in two ways. First of all, we provide intertextual links directly between individual passages. Many existing systems, such as the *Web of Science* citation index tools, can only indicate that a particular work cites another. While this is adequate for finding out which journal articles are important in a particular sub-field, it is not as useful for literary studies, where the particular portion of a poem or play being cited is important. Second, the intertextual links within Perseus are not displayed on a generated search-results page or frame but directly within the text, as the user reads it. There is no need to invoke a special search tool or search mode; citations are always available for reference.

The most obvious use for intertextual links is for commentaries. A commentary always cites the text it is commenting on, and is generally not read separately from the text it comments on. Commentaries within Perseus are linked bi-directionally with their texts. Readers can distinguish commentary links from other intertextual links by the symbol used to mark them (a superscript asterisk instead of a cross).

Intertextual citations are not limited to grammars and commentaries. Any text may cite another, and editorial footnotes to one text may also cite other texts. All of these citations generate hyperlinks to the quoted text at the point of the quotation, and also generate reference hyperlinks back from the quoted text to the referring text. Readers can see at once whether the passage they are reading has been discussed elsewhere in the digital library.

The display of these intertextual links is under the user's control. Users may choose what classes of links they will see: links from commentaries on the work they are reading, references from the body of the text of other works (including grammars and commentaries on *other* texts), references from footnotes to other works, and references from indices of other works. The default configuration, recommended for most users, is to request the first two groups: links from commentaries on the current work, if any, and from the main text of other works. Links from footnotes and indices are available for readers who want more detail.

Users may also choose how intertextual links are presented. By default, if another text actually quotes the current text, the quoted words are italicized and the link marker appears at the end of the quotation. Some readers prefer not to have the quotation in italics, but want to keep the link marker within the text. Others want no marks in the text itself; in that case, the intertextual links for quotations appear at the bottom of the page along with links to passages that mention, but do not cite, the text being read.

Perseus can make intertextual links even if the quotation in one text does not exactly match the version the user is reading. Versions might differ when one text uses a variant reading, or when the quoting author deliberately does not quote exactly, perhaps omitting words or simplifying syntax. Versions will also differ if the quoting author quotes the original Greek or Latin, but the reader has chosen an English translation. When the quotation does not appear exactly in the text presented to the reader, the intertextual link is treated in the same way as a reference to the text that does not quote it: the reference appears at the bottom of the page.

All of these implicit searches – for Lookup Tool keywords, Atlas entries, and citations in other texts – are generated at display time, not coded directly into the texts. The Lookup Tool search is generated from the same database that users can search by using the Lookup Tool directly; similarly, there are databases for Atlas entries and for citations. Each of these databases is re-built nightly. This means that when new information of whatever sort is added to the digital library, it will automatically be found by implicit relevance searches from any text.

Automatic keyword and place-name recognition do not rely on any tagging in the source text. Intertextual linking relies on correct markup of citations in the quoting text, but does not require tagging in the target text beyond the basic structural markup required to implement the canonical citation scheme for the text – books and lines for Homer's *Iliad*, books and chapters for Thucydides, and so on. There are no "anchor" points in the target text. This means that if a new text is added that quotes an existing text, the existing text will receive links from the new text as soon as the citations database is re-compiled. The existing text itself does not need to be changed.

The automatic, implicit search facilities within the Perseus Digital Library make it possible for even inexperienced readers of Greek and Latin texts to make some of the connections that classical scholars have learned to make for themselves. Readers can move easily from Homer's *Iliad* to sculptural depictions of the Trojan War on the Treasury of the Siphnians at Delphi, a group of vases showing Ajax and Achilles playing board-games, or photographs of Troy; or they can move to passages of Strabo's *Geography* or Aristotle's *Rhetoric* that quote the *Iliad*. Readers do not need to ask "are there any vases that show scenes or characters from this story?" They do not even need to know that vases *might* illustrate what they are reading. The automatic search facility marks "Achilles" as a keyword and makes a hyperlink to the Lookup Tool. A reader who does not remember who Achilles was, or who wants to know more, can follow the offered link to find everything available – including vases that show scenes from the *Iliad*.

Because the integrated, implicit searches do not distinguish among texts, art works, and geography, readers of texts are prompted to look beyond the words to the rest of ancient Greek and Latin culture. Undergraduates in elementary language courses (or graduate students frantically preparing for qualifying exams) may be tempted to focus too closely on morphology, syntax, and lexicon. The implicit search tools of the digital library always offer a wider view, enriching the experience of reading the text and discovering these cultures.

Although I have focused on examples from classical studies in this paper, the techniques described here are extensible to other domains. The Perseus Digital Library also includes early modern English literature (Shakespeare and Marlowe); scientific works from the same period; and geographic and sociological data for London, focusing on the 19th century but extending back to the early modern period. We use the same text engine and implicit search facilities for all of these collections.

References

Agosti, M., M. Melucci, F. Crestani. “Automatic Authoring and Construction of Hypermedia for Information Retrieval.” *Multimedia Systems* 3(1995), 15-24.

Berk, E., and J. Devlin. *Hypertext/Hypermedia Handbook*. New York: 1989.

Chavez, Robert F. “Generating and Reintegrating Geospatial Data.” Digital Libraries ’00 Short Papers.

Crane, Gregory. “The Perseus Project and Beyond: How Building a Digital Library Challenges the Humanities and Technology.” *D-Lib Magazine*, January 1998, http://www.dlib.org/dlib/january98/01crane.html.

————————. “Generating and Parsing Classical Greek.” *Literary and Linguistic Computing* 6(1991), 243-245.

Frakes, William B., and Ricardo Baeza-Yates, eds. *Information Retrieval: Data Structures and Algorithms*. Upper Saddle River: 1992.

Kaindl, H., S. Kramer, and P. S. Niang Diallo. “Semiautomatic Generation of Glossary Links: A Practical Solution.” *Hypertext 99: Proceedings of the 10th ACM Conference on Hypertext and Hypermedia*. New York: 1999; 3-12.

Marchionini, G. M. *Information Seeking in Electronic Environments*. Cambridge: 1995.

Salton, G., and M. J. McGill. *Introduction to Modern Information Retrieval*. New York: 1983.

Smith, David A., Jeffrey A. Rydberg-Cox, and Gregory Crane. “The Perseus Project: A Digital Library for the Humanities.” *Literary and Linguistic Computing*, forthcoming.