Expanding the Notion of Links

Steven J. DeRose

Summer Institute of Linguistics
7500 W. Camp Wisdom Road
Dallas, TX 75236

INTRODUCTION

Research in the humanities, particularly in text-oriented fields such as Classics and Religious Studies, poses particular challenges to hypertext and hypermedia systems. The complex set of primary and secondary documents form an intricate, highly interconnected network, for the representation of which hypertext is ideal. The variety and quantity of links which are needed pose challenges especially for data structures and for display and navigation tools. The specific needs arise in other contexts as well, particularly those with very large or complicated document collections.

In this paper I shall classify and discuss these needs, with illustrations from the CD-Word project at Dallas Theological Seminary,[1] the Perseus Project at Harvard University,[2] and a variety of other hypermedia systems.

ISSUES IN SUPPORTING SCHOLARLY TEXT RESEARCH

Classical and biblical documents, like other natural language texts, have *more than one structure*: (a) a logical or linguistic structure, with units such as chapters, paragraphs, sentences, etc.; and (b) a physical or layout structure, with pages, lines, etc. Although the logical structure is more important, for certain purposes the layout is also needed (see below). Both structures are largely hierarchical, but they cannot readily be reconciled into one hierarchy. This poses difficulties for many systems (such as Guide) which constrain document structures and links to a single hierarchy. Even worse, many systems do not support hierarchies *per se* at all, though one can usually build fortuitously hierarchical structures; the fact that hierarchies are a real part of text structure means that completely free-form systems such as HyperCard, NoteCards, HyperGate, and so on tend to miss an important aspect of documents.

---

[1]CD-Word gathers a range of primary and secondary documents for biblical studies into a Guide-based environment intended to facilitate the work of students, researchers, and clergy. It is being developed by Dallas Theological Seminary through private financial sponsorship, with the assistance of Owl International and Fulcrum Technologies Inc. I serve as a consultant to the CD-Word project, but the opinions expressed here are my own and do not necessarily reflect the views of the project or other staff. I would like to thank Robin Cover and Gary Simons for numerous helpful comments on this paper, and the Brown University Computing in the Humanities Users' Group, especially David Durand, Andrew Gilmartin, Elli Mylonas, and Allen Renear, for many enlightening discussions of hypertext.

[2]Perseus [Hugh88, Cran87] is a major project which gathers many classical Greek texts and some reference works, plus images of architecture, artifacts, and so on into a hypertextual database for pedagogy and research. Perseus, under the guidance of Greg Crane, Elli Mylonas, and others, has also produced a number of text analysis and search utilities which are being integrated into Perseus' hypermedia environment.

In addition to these two commonplace hierarchical structures, ancient documents often have an unbounded number of additional structures (some hierarchical, some not), which reflect the analytical decisions of exegetes and other textual scholars. Furthermore, many important texts have formal naming schemes which scholars use in order to "link" to pieces of the documents; the named pieces usually match elements of the main structures.

Ancient documents have another complexity: they exist in many *versions*, creating through repeated copying and in some cases editing. This is similar to the problem of document versions which has recently received attention. However, unlike with newly-authored documents, we do not know the sequential or genetic relationships between the extant copies of ancient texts; inducing them is itself a scholarly pursuit, for which the ability to compare and associate different versions, is needed. Low-end systems completely ignore version control, some even (as does HyperCard) foregoing "Undo." FRESS [Cata79, vanD71] introduced "Undo," but only to one level, with no permanent record. Many systems support linear "audit trails," but assume there is only one successor to each version. Nelson's Xanadu design allows for multiple successors, but appears to fail when a single successor element has multiple sources. To my knowledge, no hypermedia system yet handles the degree of multiple inheritance required for manuscript research.

The documents of interest to literary scholars also frequently exist in versions of another kind: *translations*. A translation is characterized by having much the same document structure as its original (and, hopefully, the same meaning), but little or none of the same concrete content (at the word and character level). Of course, readers often wish to see translations in co-ordination with originals. Various classes of annotations, such as part of speech labels, are similarly relevant, but tend to apply at the low levels of structure, in contradistinction to translations. CDWord is one of the few systems so far which coordinates simultaneous display of translations.

Literature in general differs from more technical material in that it requires deeper interpretive skills; whereas a technical manual aims to be very explicit, this may not be a significant goal for a novel. Because of this, the desired paths of exploration and methods of annotation cannot be defined in advance for literature (even by the author); therefore sophisticated search, retrieval, and annotation tools are required.

Because most ancient documents are in unfamiliar languages, many users need help from dictionaries and other aids, as well as tools for locating desired passages, when skimming, especially when skimming or retrieval is impractical.

A TAXONOMY OF LINKS

These corpora show that links involve much more complicated theoretical and design issues than may at first appear. This section will describe these linking needs in terms of several sorts of links that differ not only in purpose but in structure, function, and preferred means of implementation.

Figure 1 presents the taxonomy of links which I will discuss. The precise divisions could be expressed in different ways, but I suggest that this set of relationships is useful to the user as a framework for understanding the capabilities of a system, and to the designer when planning user interfaces and data structures.

On paper I can only approximate this taxonomic structure in the arrangement of the discussion, by choosing some particular order in which to describe its components. Ideally, however, I would create a hypertext which mirrored the taxonomic structure itself.

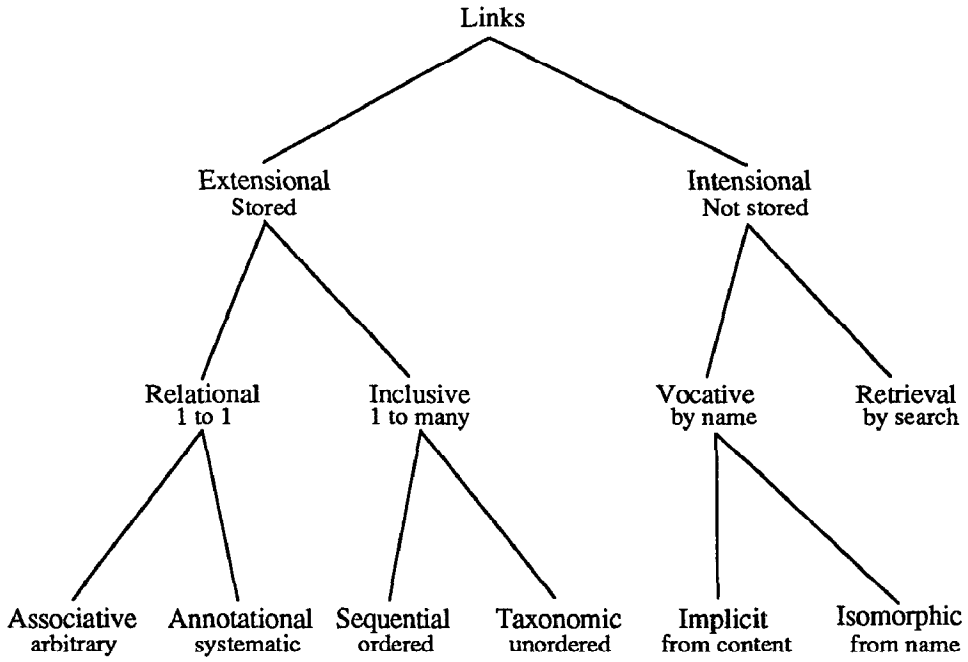

Figure 1: A taxonomy of links

1: Extensional links

Extensional links are idiosyncratic, tying various parts of the docuverse together in unpredictable ways. They must be stored individually. Extensional links have received much more attention than *intensional* links, which are discussed below as class number 2.

1A: Relational links

The first subdivision of extensional links includes *relational* links (Conklin uses the term "referential" [Conk87]). These connect single locations together. By writing "locations" I do not wish to prejudice the question of whether points, markup elements, arbitrary spans of text, or even discontiguous spans are linkable. The distinction is rather that each of the two ends of a relational link is one conceptual unit, not many.

1A1: Associative links

A relational link which is entirely unpredictable is called an *associative* link. Such links are the usual stock in trade of hypertext systems. Since they attach arbitrary pieces of documents, they cannot be replaced by retrieval algorithms, or even by unilateral creation on the part of an author. Rather, every user must be able to create them on the fly and to organize them in whatever ways seem appropriate.

Because these links serve many purposes, they are usually labelled according to type. Trigg has proposed a taxonomy that insightfully covers a large range of needs [Trig83]. However, no closed taxonomy is likely to be adequate for all future purposes, so coining new types should be possible.

There is, nevertheless, a problem with an unrestricted set of types: the problem of standardization. If users can create their own type-names, they will certainly eventually produce a situation in which one cannot effectively choose links based upon type. In a worldwide community of users there is the problem of language, but even in a single language users will find many names for the same thing, and the same name for distinct things. How, then, can one resolve the semantics of link types?

The same problem occurs in electronic mail forums, bulletin boards, etc.: systems that do not enforce a list of topics quickly bog down in inconsistency. Similar problems in library cataloging led to using standard subject taxonomies.

In the same way, new types must be possible, yet lack of standardization has its own dangers. I advocate allowing any user to add to a standard taxonomy, but only by choosing to make each new type a sub-type of one which already exists. This solution has many of the same advantages (and limitations) as type inheritance systems in object oriented programming systems.

1A2: Annotational links

An *annotational* link differs from an associative link in that one of its ends is (in principal) predictable. That is, the existence of a link from each of a class of locations is predictable, but the targets of those links are not.

Thus, part of speech marking is an annotational link, because it cannot be reliably predicted merely from a word's form; dictionary lookup is not, because the target is relatively predictable. Annotational links are very similar to isomorphic links (see below), but represent connections from portions of a text to *information about the text*, such as the presence of linguistic, thematic, or other phenomena. They also tend to originate from very low level elements (e.g., from every word), although they can originate from larger units as well.

Often an annotation is selected from a fixed set of items, such as {noun, verb, adjective, adverb, particle}. It is useful if a system can enforce such user-defined constraints, but I have seen no hypermedia system which can (but see [Simo88]).

One difficulty with annotational links is that they are likely to be attached to every word of a text. In such a case the user probably does not want to see an inline button or link marker after every word! Instead, such links should either remain invisible until requested, or perhaps be followed automatically and displayed interlinearly. In CDWord they are available through ever-present menus.

1B: Inclusion Links

Extensional links which connect one originating location to many target locations, not just one, are called *inclusion links*, and are similar to Conklin's "organizational" links [Conk87:34]. They function mainly to represent super-ordinate/sub-ordinate relationships between document elements. They are of two sub-types: sequential and taxonomic.

1B1: Sequential links

A *sequential* link has multiple, ordered target locations. *Paths* are a simple example: it should be possible to associate a path with a given location, so the path is accessible from it as a matter of course, and this feature follows immediately from viewing paths as sequential links.

However, the most important example of the sequential link is the *structure-representing* or *s-r* link, which represents those aspects of document structure commonly encoded via descriptive markup in word processing. For example, a section links to the sequence of

its sub-parts, which may include subsections, paragraphs, block quotes, emphatic elements.... Most sequential links are of this kind, and represent hierarchical structures of the text (see [Coom87]).

A major identifying feature of these structure-representing ("s-r") links is that they provide a basis for presenting the text in linear form when needed. Some documents exist that need never be formally linearized. However, millions of documents already exist in linearized form, whose authors thought carefully about that constraint when composing them, and so we ought to provide for them.

Card-oriented hypertext systems seldom support s-r links, thus avoiding real complexities of document structure by implementing an impoverished model of text. Some systems (such as Xanadu [Nels87]) treat structural relationships as merely a special case of associative links. However, s-r links deserve special treatment by hypermedia systems, for several reasons (cf [DeRo87]):

First, s-r links usually require specialized display, for example being traversed automatically in order as the user scrolls, rather than followed only on request. That is, one should not have to follow links in order to see successive verses of the Bible, speeches within an ode, or paragraphs of a chapter; rather, one should see a smooth and uninterrupted view of the *document*.

Second, s-r links express many useful and standard hierarchies. Although systems like NoteCards allow hierarchical organizations, they do not provide support for defining *specific* structures which authors consider standard, such as outline levels, chapters/sections/subsections, etc. The user who needs non-hierarchical documents is of course not constrained, since many other kinds of links are available; but by knowing about formal structures, a system can (see also [Barn88]):

1) assist in generating useful links between related structural elements (for example, collecting all section-title elements into a table of contents linked to the contents);

2) perform more effective retrieval (for example, weighing words in titles more heavily than words in running text);

3) help prevent anomalous documents (for example, those whose paragraphs contain chapters).

Third, s-r links characteristically point to sequences of other s-r links and not to arbitrary spans of text. For example, a section may be a sequence of paragraphs; it is not merely a sequence of ranges of characters. Representing a section in the latter fashion would not express that certain ranges are also the targets of successive paragraph-links, and that this is not mere chance; as Conklin points out ([Conk87:36), some elements "are much more tightly bound together than. . . nodes are to each other."

Fourth, unlike associative links, the set of s-r link types is fairly constrained. Standardization of types is important, as for associative link types. However, even with a carefully planned standard (e.g., [AAP86]), new or forgotten uses will continue to arise.

1B2: Taxonomic links

A *taxonomic* link leads to multiple target locations, but does not impose an order on them. Such links generally associate lists of properties with particular document elements. For example, one may associate examples of some literary phenomenon with commentary about it, or attach keywords indicating relevance, importance, secrecy requirements, etc. (FRESS was probably the first hypermedia system to support such information). One may create an unordered path connecting passages of interest,

otherwise like the paths created with sequential links. Another application is to connect related groups of data in a lexicon, such as cross-references between words.[1] Taxonomic links are very similar to annotational links, but tend to originate from higher-level elements.

2: Intensional links

Opposed to all the extensional link types already discussed, are the *intensional* link types. These have in common that they follow strictly from the structure and content of the documents they link, and so need not be stored one by one by the hypertext system. In other words, the destination of an intensional link is defined by some *function* that finds the desired ends, rather than being a *list* of known ends. Because of this, intensional links are in principle unidirectional, although in some cases a function may be invertible, making it possible to reverse some intensional links. They may also have multiple ends. Conklin [Conk87:35] mentions a broad notion of "keyword links" in passing, and notes they are "yet to be fully explored."

2A: Vocative links

Some intensional links are called *vocative* because they invoke a particular document element *by name*. Many document elements do not have names useful to humans. However, many also do. For example, each entry in a dictionary has its main entry word. Reference works in general are distinguished by their use of element names, but other documents also name some elements ("Chapter 3" or "Figure 27"), and many ancient texts have standard reference methods. In all these cases, certain links may be inferred, and thus need not be stored.

2A1: Implicit links

A vocative link which exists because its target element's name appears within the *content* of the source document is called an *implicit* link. The most obvious example is dictionary look-up. Dictionaries should be available from every word in every text; this clearly requires too many links to store or display explicitly, especially when many dictionaries are involved.

In the simplest case the system need merely extract a selected word and use it as an index key (or element name) of the referenced document. However, complexities arise:

1) There may be many dictionaries for a language. New ones may be added at any time, and should not require lengthy processing before being used.

2) The word form found in the text may not be the same as the index key for the dictionary; in some languages (those with complex morphology) this problem can be nearly intractable.

3) Implicit links may be used to access many other kinds of documents and sub-documents in addition to dictionaries. For example, online maps of all the world's countries ought to be accessible from any instance of a country name in any text (similar to features in Perseus [Hugh88]).

---

[1] An obvious but incorrect application of taxonomic links would be to encode formatting information. Formatting should (except perhaps in particular rare cases) be a consequence of element *types*, not of element *instances*; otherwise consistency of formatting and of format-changing is lost, and the user can be tempted to discard the role of author for that of typesetter, seldom a good use of an author's time.

4) Implicit links may attach not just to words, but to any elements of the text; if they are too frequent to have individual link markers, a clear way must be provided for choosing what element to use.

Because implicit links can are so frequent, and can go to such a wide variety of places, a means for choosing where in the docuverse implicit links are to lead is required. In CDWord, users can select a word in a text, and then choose dictionary lookup, map lookup, or other options via a menu. Since there are multiple reference works in each category, users can choose a preferred work in each category, which will then be the one accessed.

Standardized reference methods constitute a second kind of implicit link. Appearing in print, "Matt 1:10" is a link to chapter one, verse ten of Matthew's Gospel. It is an abstract link, leading to a structural element that transcends the particular manuscripts, versions, translations, and other documents that instantiate the notion "Bible." The reader reasonably expects that every such reference in any online text links to the correct verse in any document sharing certain aspects of the Bible's structure, even if the author has not created an associative link, or even marked up the reference as such. These links may be handled in much the same manner as dictionary look-ups. As Crane has pointed out, such links have an advantage over typical associative links, in that they communicate to the user some indication of their destination [Cran87:53].

Bibliographic and similar references constitute a third kind of implicit link. Internal references may be as brief as "see chapter 10." References to other works may appear in many different forms as dictated by the kind of document referenced and by personal and editorial tastes. Nevertheless such cross-references should function as implicit links. At present this is not so serious a problem, because only a small portion of literature is available in electronic form. As the docuverse becomes a reality, however, it will become crucial to standardize the syntax of bibliographic references.

Implicit links are used all the time on paper, so they will be an important part of what users desire for some time to come. As already noted, references found in printed originals pose the particular problem that they commonly point to page numbers. There will also be many occasions for making references from the docuverse out to paper documents, and so online document systems must support page references.[1]

Implicit links, finally, may come directly to the user's mind. A user should be able to request a particular dictionary entry, Bible verse, or journal article by name at any time. Most hypermedia systems allow this in some form: some require an arcane name, such as a card id; some (e.g., FRESS) have provided explicit naming operators and an index space; in cases where conventional names exist, they should be directly usable as well.[2]

2A2: Isomorphic links

A vocative link which exists because its target element's name appears as an element *name* in the source document (rather than as content) is called an *isomorphic* link. Isomorphic links are most useful in cases where different documents share much or all of their logical structure. They define the correspondences among large structures of elements, usually entire documents.

---

[1]Perhaps even more difficult is the problem of referring to elements of an online hypertext from paper, since pages are not meaningful, and element identifiers do not seem intuitive.

[2]This is harder than it may seem; for example, verses of the Psalms are numbered differently in Greek and Hebrew versions, yet must be interconnected.

I have already described the many documents that share the element structure of the Bible; documents which exist in several versions sharing most of their structure pose a similar problem. A hypertext system that includes such documents should provide for connecting the corresponding elements, so that users can compare and relate them.

A distinguishing characteristic of isomorphic links is that they tie together like-named (not "like-positioned") document elements. Despite superficial variations there exists an abstract (meta-) document we call the "Bible," represented by thousands of concrete manuscripts, editions, and translations. These documents share a structure of abstract elements. By providing the ability to name document elements similarly despite their existing in diverse documents, the needed level of abstraction can be achieved.

In reading meta-documents, the most common need involves viewing simultaneously several concrete instances of a particular text phenomenon. Parallel, synchronized windows or panes are the usual solution. For documents with entirely commensurable element names this solution is relatively easy, but there are usually deviations from perfect isomorphism. For example, certain elements may be missing in some versions, or re-ordered; numbered groups of elements may be divided or counted differently; elements may correspond to higher or lower level elements in other versions. Also, corresponding elements may differ drastically in size, in which case it is important for a hypertext system to align and move text intuitively.

2B: Retrieval links

Retrieval links are very similar to vocative links. However, where vocative links find their target by a formal name, retrieval links find their target by its content, or perhaps by both name, structure, and content. A retrieval link invokes a process to search a portion of the docuverse for something. The process may be of arbitrary complexity, and may in principle involve name-based, structure-based, and content-based decisions. Whereas implicit links are defined globally (for example, it can be a system universal that words are linked to dictionary entries), retrieval links originate only at particular locations.

The list of locations found by the process associated with a retrieval link constitutes the destination; thus the destination may change over time. So long as the universe of potential link ends does not change, and so long as the list is known to be complete, a list and a retrieval produce the same result; but the retrieval link survives changes more effectively.

Retrieval links could be subdivided into *nominal* links, which search the structure of named elements in their target document(s), and *content-based* links, which search the content. However, there will be cases in which a single retrieval link must refer to both types of information, so I do not think this distinction is salient.

CONCLUSIONS

Just to support the features of typical paper versions of the Bible, we must include hundreds of thousands of links. The large number of links required would, if kept explicit, severely tax the capabilities of any hypertext system. Thus the dense and complex interconnections of the biblical studies materials clearly demonstrate the need for more sophisticated linking methods in hypermedia. A fundamental requirement is support for intensional linkage, where an unbounded set of links is supported by indexing and retrieval rather than by exhaustive cataloging. Moreover, a wide range of link kinds should be supported, most of which do not readily fit into the usual associative linking paradigm.

Biblical, classical, and literary scholars regularly require such complex tools, and have devised means to the same ends on paper. Designers of hypermedia systems would do

well to take these problems into account, for they will probably arise in all fields, differing only in priority and guise.

BIBLIOGRAPHY

[AAP86] American Association of Publishers. 1986. *Reference Manual on Electronic Manuscript Preparation and Markup.* Washington, DC: AAP Electronic Manuscript Project.

[Barn88] Barnard, David T., Cheryl A. Fraser, and George M. Logan. "Generalized Markup for Literary Texts." *Literary and Linguistic Computing* 3 (1): 26-31.

[Cata79] Catano, J. 1979. "Poetry and Computers: Experimenting with Communal Text." *Computers and the Humanities* 13(9): 269-275.

[Conk87] Conklin, Jeff. 1987. "Hypertext: An introduction and Survey." *IEEE Computer* 20 (9): 17-41.

[Coom87] Coombs, James H., Allen H. Renear, and Steven J. DeRose. 1987. "Markup Systems and the Future of Scholarly Text Processing." *Communications of the Association for Computing Machinery* 30 (11): 933-947.

[Cran87] Crane, Gregory. "From the Old to the New: Integrating Hypertext into Traditional Scholarship." In *Proceedings of Hypertext '87.* Chapel Hill: Department of Computer Science, University of North Carolina.

[DeRo87] DeRose, Steven J. 1987. "Hypertext and Scholarship in the Humanities." Position paper in *Proceedings of Hypertext '87.* Chapel Hill: Department of Computer Science, University of North Carolina.

[Hugh88] Hughes, John. 1988. "Studying Ancient Greek Civilization Interactively—The Perseus Project." *Bits and Bytes Review* 2 (1): 1-12.

[Nels87] Nelson, Theodore H. 1987. *Literary Machines.* Edition 87.1. South Bend, Indiana: The Distributors.

[Simo88a] Simons, Gary F. 1988. "The Computational Complexity of Writing Systems." In *Proceedings of the Fifteenth LACUS Forum.* Lake Bluff, Illinois: Linguistic Association of Canada and the United States.

[Simo88b] Simons, Gary F. and John V. Thomson. 1988. *How to Use* IT: *Interlinear Text Processing on the Macintosh.* Edmonds, Washington: Linguist's Software.

[Trig83] Trigg, Randall H. 1983. "A Network-based Approach to Text Handling for the On-Line Community." Ph.D. dissertation, University of Maryland.

[vanD71] van Dam, Andries and David E. Rice. 1971. "On-line Text Editing: A Survey." Computing Surveys 3 (3): 93-114.

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
© 1989 ACM 089791-339-6/89/0011/0257 $1.50