Services for a Customizable Authority Linking Environment

David Mimno  
Tufts University  
mimno@perseus.tufts.edu

Mark Patton  
Digital Knowledge Center  
Johns Hopkins University  
mpatton@jhu.edu

ABSTRACT
The SCALE – Services for a Customizable Authority Linking Environment – project is developing tools to help integrate collections within the National Science Foundation’s National Science Digital Library (NSDL) through terminological linking. These tools will transparently provide reading support to NSDL collections by automatically linking phrases in the content of one collection to relevant material in another collection. This will enhance the user experience in the NSDL and encourage greater integration of collections by increasing the value of each collection to other collections.

Categories and Subject Descriptors
H.5.4 [Information Interfaces and Presentation]: Hypertext/Hypermedia; I.7.2 [Document and Text Processing]: Document Preparation—Hypertext/hypermedia

General Terms
Design, Human Factors

Keywords
Digital Library, NSDL, Reading support

1. INTRODUCTION
The National Science Digital Library is an ambitious project that seeks to meld disparate collections and services into a coherent whole. Readers of a document in one collection should be able to move seamlessly to relevant supporting material in another collection. SCALE will provide tools for collection maintainers to create and update these document links. These tools will support harvesting and aggregating technical terms and associated authoritative reference materials; matching instances of a terms to appropriate references; and displaying documents with terms linked to these references. In this way, SCALE will help the NSDL integrate collections and better serve users.

2. ARCHITECTURE
SCALE models reference materials as lists of terms associated with internet URIs. These lists are transported in an XML format to a database where they are aggregated and indexed. The database is maintained by automatically harvesting lists from collections. Collections take responsibility for creating and maintaining harvestable lists derived from their own content.

Our work to date has focused on the reference linking process itself. That process consists of two steps: finding references by matching a document against available reference terms and displaying the document with associated references in a usable way. Most document types are converted to HTML for processing and display.

In order to determine the most appropriate reference links for a given document, we extract the text of the document and match the words against a database of term-reference links. We then filter any resulting term matches to find the most relevant links. Relevance is determined by a combination of statistical methods and user preferences. For example, it may be possible to guess whether a reference to “minerals” should be linked to a reference entry about nutrition or about rocks, depending on the context of the page. In another case, a user may have identified herself as an undergraduate, in which case we would filter out reference articles aimed at middle-schoolers. Additionally, users will have a high degree of control over both the matching process and the presentation.

SCALE displays a processed document to the user by incorporating reference links directly into the document. HTML, Javascript, and style sheets give a lot of flexibility over the resulting look and feel, but browser support for standards is not uniform. To take advantage of that flexibility and still accommodate a variety of users, a number of customizable methods are available. Each method must at least indicate which terms are matched, show associated references, and link to referenced content. Other important considerations include not harming the original presentation and gracefully handling terms with more than one reference. Display methods we have implemented include highlighting matched terms and displaying links on mouse over, directly linking matched terms to references, or displaying matched terms linked to references in a separate area of the interface.

3. ACKNOWLEDGMENTS
The work of this project is funded by the National Science Foundation Division of Undergraduate Education NSDL Program (awards DUE 0226234 and DUE 0226304).