Open Identification and Linking of the Four Ws

Ryan Shaw
University of California, Berkeley, USA
ryanshaw@ischool.berkeley.edu

Michael Buckland
University of California, Berkeley, USA
buckland@ischool.berkeley.edu

Platforms for social computing connect users via shared references to people with whom they have relationships, events attended, places lived in or traveled to, and topics such as favorite books or movies. Since free text is insufficient for expressing such references precisely and unambiguously, many social computing platforms coin identifiers for topics, places, events, and people and provide interfaces for finding and selecting these identifiers from controlled lists. Using these interfaces, users collaboratively construct a web of links among entities.

This model needn't be limited to social networking sites. Understanding an item in a digital library or museum requires context: information about the topics, places, events, and people to which the item is related. Students, journalists and investigators traditionally discover this kind of context by asking "the four Ws": what, where, when and who. The DCMI Kernel Metadata Community has recognized the four Ws as fundamental elements of descriptions (Kunze & Turner, 2007). Making better use of metadata to answer these questions via links to appropriate contextual resources has been our focus in a series of research projects over the past few years. Currently we are building a system for enabling readers of any text to relate any topic, place, event or person mentioned in the text to the best explanatory resources available. This system is being developed with two different corpora: a diverse variety of biographical texts characterized by very rich and dense mentions of people, events, places and activities, and a large collection of newly-scanned books, journals and manuscripts relating to Irish culture and history.

Like a social computing platform, our system consists of tools for referring to topics, places, events or people, disambiguating these references by linking them to unique identifiers, and using the disambiguated references to provide useful information in context and to link to related resources. Yet current social computing platforms, while usually amenable to importing and exporting data, tend to mint proprietary identifiers and expect links to be traversed using their own interfaces. We take a different approach, using identifiers from both established and emerging naming authorities, representing relationships using standardized metadata vocabularies, and publishing those representations using standard protocols so that links can be stored and traversed anywhere. Central to our strategy is to move from appearances in a text to naming authorities to the the construction of links for searching or querying trusted resources.

Using identifiers from naming authorities, rather than literal values (as in the DCMI Kernel) or keys from a proprietary database, makes it more likely that links constructed using our system will continue to be useful in the future. WorldCat Identities URIs (http://worldcat.org/identities/) linked to Library of Congress and Deutsche Nationalbibliothek authority files for persons and organizations and Geonames (http://geonames.org/) URIs for places are stable identifiers attached to a wealth of useful metadata. Yet no naming authority can be totally comprehensive, so our system can be extended to use new sources of identifiers as needed. For example, we are experimenting with using Freebase (http://freebase.com/) URIs to identify historical events, for which no established naming authority currently exists.

Stable identifiers (URIs), standardized hyperlinked data formats (XML), and uniform publishing protocols (HTTP) are key ingredients of the web's open architecture. Our system provides an example of how this open architecture can be exploited to build flexible and useful tools for connecting resources via shared references to topics, places, events, and people.

References

Kunze, John A. and Adrian Turner. (2007). *Kernel Metadata and Electronic Resource Citations (ERCs)*. Retrieved June 13, 2008, from http://www.cdlib.org/inside/diglib/ark/ercspec.html.