

Translating (human-scale) information use into information integration

Lois M. L. Delcambre and David Archer
Computer Science Department
Portland State University
{lmd, darcher}@cs.pdx.edu

1. Superimposed information: capturing and reusing human attention

With the simple act of copy-and-paste, a user can easily extract excerpts from a myriad of existing information sources and then combine and elaborate the extracted information in an unlimited number of new contexts. But we can do much more by automatically capturing the provenance of the excerpts and allowing the users to revisit the original information, in context. Our work with our colleagues over the past decade has been developing the notion of *superimposed information* [3, 4, 6, 7, 8, 9] where the user can easily create a mark to any selected bit of information. A *mark* is an encapsulated address, with associated metadata, that can be stored in a repository and easily referenced and used from a wide variety of what we call *superimposed applications*. The key feature of a mark is that the user can always return to the original application, with the original selection highlighted, in its original context. Our work to date has required minimal support from existing applications – namely the ability to *create a mark* and the ability to *go to mark*.

We have built various superimposed applications that use marks including:

- a simple scratchpad tool (called SLIMPad [4], RIDPad, and now Sidepad [5]) that allows the user to place and arrange marks (with a title and an optional comment) into groups;
- a superimposed schematic tool that allows the user to introduce an entity-relationship-style model over a set of marks – to support browsing through the underlying set of base documents [2];
- a mashup tool that assists with the selection and preparation of descriptive data for various locations on a map (based in part, on geocoding an address) [10];
- two video composition tools (SuperMix and SIMPEL [11]); and
- an interactive, superimposed, strand map-based [1] lecture that allows students to browse to mini-lectures attached to learning objectives in the strand map.

We have also defined a namespace to represent marks as a URL so that marks can be used in a wide variety of existing tools, including a concept map tool [12]. We have explored the use of our superimposed tools by students and researchers preparing to write a paper, faculty preparing a lecture, physicians preparing notes for a colleague who is handling her hospital rounds over the weekend, open source intelligence analysts who are assembling evidence to support or refute a hypothesis, lawyers working on patent claims, and, most recently, corporate managers and decision makers, e.g., evaluating employees or managing projects based on earned value. In general, we say that superimposed tools allow us to *capture and reuse human attention*, i.e., the human effort expended in selecting, elaborating, and rearranging bits of information. We see this human attention as a valuable form of metadata that can be captured as the user completes the detailed task at hand.

2. An opportunity: capturing and reusing (implicit) information integration decisions

We observe that users are often resolving data integration conflicts as they use and extract data. For example a corporate manager may easily combine information about Robert Williams in a human resources database with information about Bob Williams in the project management

system and (implicitly) perform the “entity resolution” step for this one person, across these two information sources. In a similar way, if a manager used the *salary* field from the human resources database as the source for salary for one person and the *hourly rate* field from the project management system as the source for salary for another person, then they have (implicitly) decided that these two schema elements (*salary* and *hourly rate*) match.

In general, we are working on extending our work on superimposed information to make these kinds of information integration decisions explicit. We seek to define appropriate data structures to represent this kind of information; we seek to make it very easy to capture this kind of information (as the human user is performing their usual tasks); and we seek to make it very easy to reuse this kind of information when the user needs to perform the same task or a similar task at a later time. We are also exploring the use of what we call *injection marks* to go beyond simple revisiting of marked information to extracting and repackaging marked information in new information products.

Classical database integration has led to a deep understanding of the steps involved in integrating schemas (sometimes called “model” matching), in integrating data (including entity resolution), and in supporting queries and updates through the integrated, global schema. And the use of ontologies to link data in multiple information sources is invaluable. Our work is synergistic and complementary to this type of work. We observe that ad-hoc information integration tasks are performed all the time by users who have no training in data management. We seek to assist with and exploit the innumerable settings where users are already doing ad-hoc integration, where a suitable schema or ontology may not yet exist.

3. The potential benefit

If we succeed at capturing information integration decisions¹ as they occur, then the benefit potentially goes far beyond assisting users in the task at hand. If such information integration decisions made by expert users can be combined, then this may be very useful for automatic or semi-automatic methods, e.g., as training data, to drive automatic, task-driven ontology creation or ontology class selection, potentially bringing the benefits of Semantic Web technology to a broad class of inexpert users.

References

1. AAAS. (2001). "Atlas of Science Literacy, Project 2061." American Association for the Advancement of Science, National Science Teachers Association, Co-Publisher, Washington, DC.
2. Shawn Bowers, Lois M. L. Delcambre, David Maier: Superimposed Schematics: Introducing E-R Structure for In-Situ Information Selections. ER 2002: 90-104
3. Delcambre, L. and Maier, D. (1999). "Models for Superimposed Information." *Lecture Notes in Computer Science* **1727**: 264 - 280.
4. Maier, D. and Delcambre, L. (1999). Superimposed Information for the Internet. *In Proceedings of the WebDB Workshop*: 1-9.
5. Murthy, S. (2005). "Sidepad User Guide"
<http://datalab.cs.pdx.edu/sparce/apps/Sidepad/userguide/index.html>.
6. Murthy, S. and Maier, D. (2003). SPARCE: Superimposed Pluggable Architecture for Contexts and Excerpts. OGI CSE, Technical Report #CSE-03-010, <ftp://ftp.cse.ogi.edu/pub/tech-reports/2003/03-010.pdf>.

¹ We have a working prototype of our “marionette” system that supports injection marks as well as tracks entity- and schema-resolution decisions, using sample data from a corporate management environment.

7. Murthy, S., Maier, D. and Delcambre, L. (2004a). Querying Bi-level Information. *In Proceedings of the the Seventh International Workshop on the Web and Databases (WebDB 2004)*, Paris, France: 7-12.
8. Murthy, S., Maier, D. and Delcambre, L. (2005). "Distribution Alternatives for Superimposed Information Services in Digital Libraries." *Lecture Notes in Computer Science* **3664**(Aug 2005): 96.
9. Murthy, S., Maier, D., Delcambre, L. and Bowers, S. (2004b). Putting Integrated Information into Context: Superimposing Conceptual Models with SPARCE. *In Proceedings of the First Asia-Pacific Conference of Conceptual Modeling*, Denedin, New Zealand: 71-80.
10. Sudarshan Murthy, David Maier, Lois Delcambre , Mash-o-matic, To appear in: Proceedings of the Sixth ACM Symposium on Document Engineering (DocEng 2006); 2006; Oct. 10-13; Amsterdam, Netherlands.
11. Murthy, U., Ahuja, K., Murthy, S. and Fox, E. A. (2006). "SIMPEL: A Superimposed Multimedia Presentation Editor and Player." to appear as a demo description in Proceedings of the Joint Conference on Digital Libraries 2006, Chapel Hill, USA, <http://si.dlib.vt.edu/publications/de214-murthy.pdf>.
12. Murthy, U., Fox, E. A. and Delcambre, L. M. (2006). "Enhancing Concept Mapping Tools Below and Above to Facilitate the Use of Superimposed Information." submitted to the Second International Conference on Concept Mapping, San Jose, Costa Rica Sept. 5-8, 2006, <http://si.dlib.vt.edu/publications/SlandCMapsPaper2006.pdf>.