

Superimposed Information for the Internet

by

David Maier and Lois Delcambre
Computer Science and Engineering Department
Oregon Graduate Institute
P.O. Box 91000
Portland, Oregon 97291-1000
maier@cse.ogi.edu imd@cse.ogi.edu

1 Introduction

It has existed for several millennia, in the form of commentaries on religious books, law and literature. We see it today in concordances, citation indexes and genome maps. You probably have created some of it, as a bookmark file in your web browser. The “it” we refer to is *superimposed information*: data “placed over” existing information sources to help organize, access, connect and reuse information elements in those sources.

Digital forms of superimposed information are now common, especially in the milieu of the World-Wide Web. Some of it has an individual focus, such as the aforementioned bookmark files, or link pages on a person’s web site. Other forms are collective efforts and meant for wide audiences, for example, web guides such as Yahoo. Examples exist outside of the web, such as repositories for software projects that link together specifications, schemas, databases and program components, and directories of scientific information that allow location of datasets via associated metadata.

While superimposed information is nothing new, we expect its creation and use will increase markedly in the coming decade, for a variety of reasons. More and more kinds of information are being converted to digital form and placed on line. Moreover, that information is often addressable at a finer granularity than its hardcopy analogs: pages and paragraphs rather than entire books and articles; frames and scenes rather than whole movies and videos. The huge volumes of on-line information demand alternative groupings and organizations of information elements to make it usable and comprehensible by individuals and special interest communities. The low cost with which information can be placed on the Internet means that much of it is inaccurate or of questionable value, creating the need for annotation and evaluations by others. Finally, emerging standards such as RDF, XLink and Topic Navigation Maps will facilitate the creation and exchange of superimposed information.

Why do we think superimposed information is an important topic for future research? At the most basic level, it is an interesting phenomenon with deep historical roots that is being profoundly affected by the digital age. More pragmatically, we think a better understanding of the connection between the structure of superimposed information and the capabilities it supports will have value in designing new superimposed information models and accompanying technology. That understanding can also influence the form and function of standard for the representation and dissemination of superimposed information. Defining the common architectural elements of superimposed information systems can be the basis for frameworks and tools for more easily building such systems. Such architectures can also help understand what is required and desired of an underlying information space to better support creation and maintenance of superimposed information over it. Finally, managing superimposed information presents interesting challenges for traditional data management technology, such as handling

dynamically discovered data types, combining structured and semi-structured information, and bi-level query processing.

2 Superimposed Information: What and Why

We intend the term “superimposed information” to be construed very broadly, to denote information overlaid on other, existing *base* information for a variety of purposes. The base information might be very limited in scope (a single book) or quite extensive (all web pages). We tend to think of the base information as retaining its original form, while the superimposed information refers to it. For example, a concordance (directory of word occurrences) for Shakespeare’s plays exists as a separate document, with references to lines in the plays. (Those references might take the form of *Play, Act, Scene, Line Number*, where the line numbering is relative to a standard edition.) Thus, in general, we regard superimposed information as different from a restructuring or alternative view of the base information. Rather, superimposed information leaves the base information as is, and adds supplementary information over it (though it may be that all or part of the superimposed information can be derived from the base information). Thus, it is possible to have several different kinds of superimposed information over the same base information. For example, we can have both a concordance and a commentary for the Pentateuch.

As noted in the introduction, superimposed information greatly predates the advent of digital information. We will argue a bit later why the digitization of information has expanded the use of superimposed information, and promised to expand its role even further in the future. However, for our initial discussion, we will draw many of our examples from “hard copy” forms of superimposed information. Also note that it makes sense to have digital superimposed information over physical base information (e.g., an on-line catalog of the works of a painter) and even vice versa (a book listing and describing web sites).

2.1 Uses of Superimposed Information

Superimposed information has been produced in the past for a variety of purposes. We describe a few here.

- **Location:** The superimposed information may be intended to help located portions of the base information. Examples are the index and table of contents of a book, a concordance, a library card catalog and the indices maintained by web search engines.
- **Annotation:** The superimposed information might serve to explicate, evaluate, correct or refute the base information. Examples are commentaries on theological works, a collection of movie reviews, or an errata sheet for a textbook.
- **Connection and Comparison:** Superimposed information can be used to connect and compare multiple information elements in the base layer. An example is the Science Citation Index, which connects publications to journal articles that cite those publications. (The Google System [BP98] is kind of a web analog in which web links play the role of citations.) An example of comparative superimposed information might be the catalog of an art exhibit that looks at the borrowings and allusions between specific paintings of Picasso and Matisse.

- **Classification and Organization:** Superimposed information can serve to group and categorize elements in the base layer. An example here is the *ACM Guide to Computing Literature*, which categorizes computer science publications in a number of ways.

2.2 Why Now?

While various forms of computerized information have been around for half a century, and many computer-based forms of superimposed information already exist, we expect to see a rapid growth in digital forms of superimposed information in the near future. We discuss here some of the factors influencing this trend.

- A. **Broadening Range of Digital Information.** Both the amount and kinds of digital information are expanding rapidly. Existing documents of many kinds are being brought on line, and most forms of information—documents, images, movies, maps, recordings—now have digital formats. We will discuss below that digital information tends to be easier to reference than “hard copy” versions, so the expanding range of digital information provides more and more potential base information over which to erect superimposed information.
- B. **Accessibility.** As information appears in digital form, most of it is made accessible remotely, mainly over the Internet. Thus, a reference to a base information element is generally easily resolved into that element itself. Having a URL for a piece of information generally means the information can be obtained without leaving your desk. Contrast this situation to that of just a decade ago. While annotated bibliography might help you identify articles of interest, obtaining those articles could involve a great deal of time and labor: trips to the library, inter-library loan requests, borrowing copies of proceedings, letters to authors, searching filing cabinets. Since accessibility is now so much improved, the utility and benefit of superimposed information is greatly enhanced, making it more worthwhile to produce.
- C. **Better Addressability.** With “physical” base information, the addressing schemes available for referencing it were diverse and not very precise. Books, compositions, paintings and films were referred to in different ways. Furthermore, the granularity of reference was quite coarse, such as a whole book, or maybe a chapter thereof. Any finer degree of addressing, such as a page number, would be dependent on a particular edition or even printing of the book. Only a few works of great interest, such as the Bible and Shakespeare’s plays, have standard addressing modes that are generally recognized (*Book, Chapter, Verse; Play, Act, Scene, Line*). On the Internet, global addressing schemes (URLs, URNs) have emerged that can address virtually any kind of information, often at a much finer granularity (page, paragraph) than before. Proposed standards, such as XPointer [XP98], mean even finer levels of addressing will be available (XML elements or even individual words). This improved addressing supports information by greatly increasing the range of things that can be easily addressed, both in number and in detail.
- D. **Ubiquity of Protocols.** A small number of protocols for information access that are universally available have emerged on the Internet (http, FTP, WAIS, IIOP). That fact makes writing tools that process superimposed information relatively easy to write, at least as far as network access is concerned. In the past, while a large amount of information existed in computerized form, and on sites with network connectivity, there were no widely standardized ways to reach this information. While I might have been willing to give the world access to information in my local database, the protocol to do so was likely specific to my site or DBMS, and a tool to access it wouldn’t be very generic.

- E. **Standard Base Types.** Along with standardization of protocols, a set of base types (really formats) for digital information have emerged that are widely understood: .html, .wav, .gif, .mov. Having some agreement on representation of values of different types means that information elements of those types are separable from their underlying information sources. That is, that they can be rendered for users in an environment other than the one in which they are stored. This capability makes tools for manipulating superimposed information both simpler to write and more portable.
- F. **Cheap Server Resources.** CPU and storage costs have dropped to the point where for the most part it is not worth the bother to charge others for their use. Thus, many organizations put up information servers that will expand large numbers of cycles and disk accesses for the benefit of outsiders, a situation that would have been largely unthinkable even a decade ago. Thus, a superimposed information system can count on other sites to perform tasks, such as address resolution, on its behalf. That situation simplifies the construction of superimposed information systems.
- G. **Semi-structured Data:** The emergence of new technology for the representation and exchange of semi-structured data, such as XML [XML98], bodes well for the management of superimposed information. It allows superimposed information to be flexible, largely self-describing, and easily extensible.
- H. **Emerging Standards:** That standards for superimposed information are starting to appear is perhaps more a reflection that its importance is already recognized than a predictor of its growth in the future. Some examples are the Resource Description Framework [RDF99], XLink [XL98], and Topic Navigation Maps [TNM99].

3 Limitations of Database Approaches

Certainly restructuring and reuse of information has long been addressed in the domain of databases, for example, with view mechanisms and data integration. However, we want to point out that these traditional approaches don't capture all aspects of superimposed information and that they are not necessarily well suited for information sources on the Internet.

Probably the biggest difference between superimposed information and database-style views is that superimposed information, in general, may contain data that is not explicitly present in the underlying sources. While some forms of superimposed information, such as that maintained by web search engines, can be derived automatically from the underlying information space, other forms, such as web guides, hold "value-added" knowledge not necessarily contained in the base information. View and data integration mechanisms, in contrast, typically only contain data that is already present in the underlying sources, albeit possibly filtered, regrouped and restructured.

A second divergence is that database approaches typically require a schema for the information sources involved, and often assume a commonality of data structuring. For example, much data integration work deals with combining sources that are all relational databases. In contrast, on the web, many information sources over which we would like to superimpose information are unstructured or semi-structured, with no explicit schema. Further, the forms of information are widely diverse. New types of data are showing up online every day, which makes the "schema-first" paradigm of DBMSs a liability. A superimposed information system may want to handle new datatypes as they are discovered, without having to anticipate and pre-define schemas that are able to hold these types.

A further small point is that database systems have typically provided support and services only over data that they directly control. They aren't set up to deal with information "outside the box." Superimposed information, on the other hand, inherently involves connecting with data controlled outside the local system. (We note that DBMS features are showing up for managing external data, such as DB2's FileLinks.)

Finally, we note that traditional data integration solutions have been rather heavyweight undertakings. They typically require a good deal of up-front work--semantic analysis, schema integration, and query mapping—on a source-by-source basis. Such approaches don't seem tenable for connecting pieces of information residing in hundreds or thousands of different sources. These approaches are also costly. They may be affordable for projects with a company-wide or community-wide benefit, but are hard to justify for providing capabilities to an individual or small group.

We believe that database systems and DBMS-related integration technologies will play a role in the management of superimposed information. However, there is clearly a need to explore other parts of the data management design space for more flexible and lighter-weight technologies as well.

4 Conceptual Architecture for Superimposed Information

A superimposed information space is an information space with a base layer and a superimposed layer. The base layer consists of information sources where each information source may be organized according to a typed structure and may have various capabilities for manipulating and presenting information from the base layer. The superimposed layer is also organized according to a typed structure (of varying levels of sophistication).

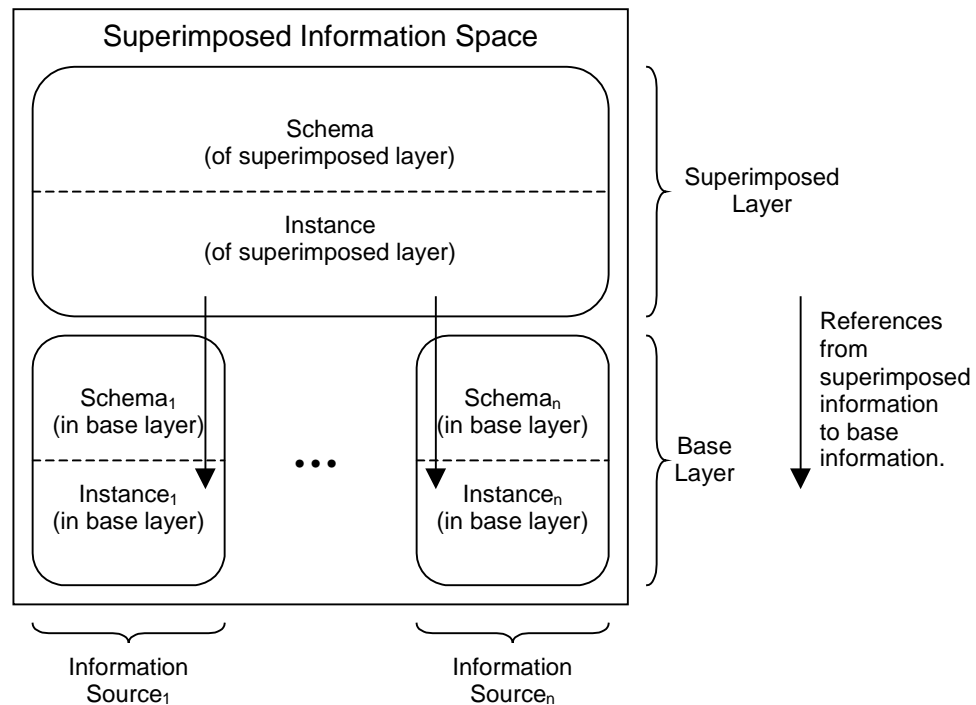


Figure 1: The Conceptual Architecture of a Superimposed Information Space

One of the key features of a superimposed information space is the ability of the superimposed information to reference information in the base layer. Note that we view the superimposed and base layers as being conceptually distinct; they may or may not be physically distinct. It is also possible that there are many levels of superimposed layers. We focus here on describing how one layer superimposes over another.

The conceptual architecture for a superimposed information space is shown in Figure 1. Each information source, shown in the bottom half of the superimposed information space, can be viewed as a pre-existing collection of information. An information source may have a simple structure such as a collection of HTML pages or a more complex structure, e.g., specified by an XML DTD or a relational database schema. We show in Figure 1 the structure of the information source; as schema_i. The actual information (that conforms to the structure) is shown as instance_i in the figure. Each information source may have a different structure.

The superimposed layer is shown on the top half of Figure 1. The superimposed layer also consists of a schema and an instance, where the schema describes the permissible structure of the information in the superimposed layer. Said another way, we view the superimposed layer as a mechanism to highlight, interconnect, and elaborate information in the base layer. Note that the model used in the superimposed layer (and hence the schema of the superimposed layer) can vary greatly in sophistication. We examine issues concerning models for superimposed information in a companion paper [DM99].

The superimposed information may also be expressed in a model that is not used by any of the information sources in the base layer. One example of a superimposed model that is distinct from the base layer model is the Resource Description Framework (RDF), which associates metadata of various kinds with existing resources [RDF99]. The RDF model consists of resources where some resources in RDF correspond to web pages in the base layer. The basic model for RDF is a triple that associates a particular property and value to a resource. This model of resource, property name, property value triples is clearly distinct from the model of the base layer, e.g., HTML or XML or other models that supports Uniform Resource Identifiers.

5 An Example Superimposed Information System: Structured Maps

Structured Maps [DMRA98] were inspired by the Topic Navigation Map [TNM99]. Topic Navigation Maps have been developed in the SGML community to provide an integrated table of contents, glossary, and index for one or more SGML documents (e.g., books). Structured Maps introduce an entity-relationship model in the superimposed layer. Figure 2 illustrates a Structured Map to track artists and painting.

Each entity type may have one or more facets, intended to hold references to information elements in the base layer. That is, facets are intended to hold marks. In Figure 2, we see two entity types and one relationship type for Artist, Painting and Painted-by, respectively. The Artist entity type has a facet to reference information elements in the base layer that contain biographical information about the artist. Painting, on the other hand, has two facets: one for photographs of the painting and one for critiques of the painting. Thus each entity is elaborated through typed collections of marks on the facets. The Painted-by relationship serves to connect Artist instances with Painting instances.

Structured Maps adopted the basic Topic Navigation Map model but in a database context. In particular, we used a relational database schema to model and store the superimposed layer. This allows the superimposed layer to be queried using SQL, a feature not currently included in the

Topic Navigation Map Standard. For Structured Maps we implemented two versions of referencing the base information. In our first prototype, we superimposed our relational layer over SGML documents and thus we used SGML ids as values for addresses. In our second prototype, we superimposed our relational layer over HTML and used href references. The second prototype provides an integrated web browser for both the superimposed and the base layers.

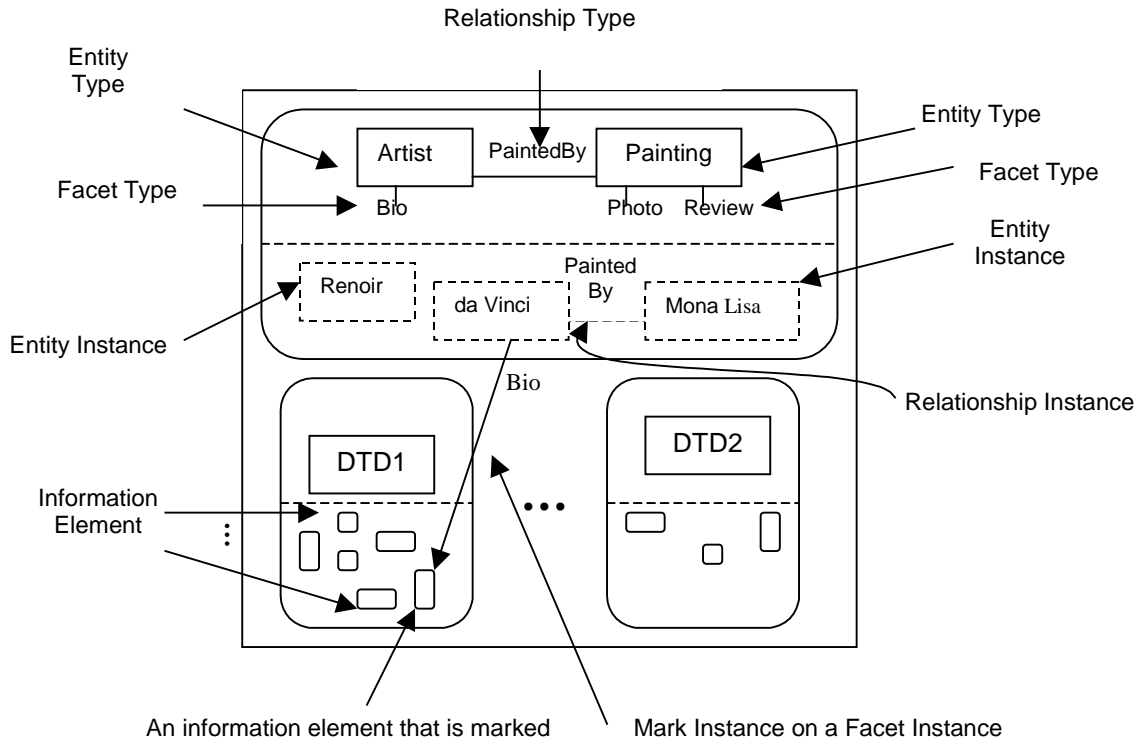


Figure 2: Example Structured Map

6 Open Research Issues

We think we have only begun to scratch the surface in our investigations of superimposed information. At this point we are producing questions at a faster rate than answers. We list some of them here.

1. Are there other forms of superimposed information in use that are not described well by the notions of information elements, marks and links as presented here? How do collections and classification fit into the picture?
2. What is the connection between features of a superimposed information model and the capabilities it supports? For example, what model features are necessary for query? For navigation? For classification-based search?
3. How does the form of superimposed information affect the effort required for its construction and maintenance? Are there some forms that are easier to elicit from users or produce using tools? Are some forms more robust in the face of updates in the underlying information space? What forms of superimposed information map easily into current

information management tools, such as relational and object-oriented databases or XML managers?

4. What are the challenges when the superimposed information has a substantially different model from the underlying information sources? We are particularly interested in highly structured superimposed information (e.g., relational tables) over lightly structured base data (e.g., web pages) and vice versa.
5. With a bi-level information structure (base and superimposed information), one might want tools that deal simultaneously with both levels. One example would be a browser from which you can hop up from the base layer to the superimposed layer, traverse connections in the superimposed information, then drop back down into the base layer. Another is a query processor that allows queries that span superimposed and base layers. The latter poses some interesting linguistic and semantic challenges when the models at the two levels are different.
6. How important is it for the superimposed layer to explicitly identify or delimit its universe of discourse in the base layer? That is, should the superimposed layer define the range of base information elements that it refers to, or may potentially refer to? (Note that it is useful to know that some range of base information was considered but no that relevant elements found.) If the universe of discourse is described, what are suitable notations for it?
7. We suspect that fusing collections of superimposed information might be substantially easier than full integration of the underlying information sources, especially if the superimposed model is simple. However, this suspicion is just an intuition at this point and needs further investigation.
8. Do the same models and techniques for superimposed data handle superimposed metadata? We are particularly interested in the notion of a “schema-later” database, where commonalities in structure or semantics are observed after some amount of data has been collected, and a schema is then “retrofitted” to the data.
9. What are the major variations in the superimposed information architecture we have sketched? One interesting case we have identified is when superimposed and base levels share a common model (both relational, both XML, etc.). With a common model, the levels could be *segregated* (stored and controlled as separate spaces) or *commingled* (managed in the same space). An obvious extension to our architecture is “super-superimposed information”: superimposed information over other superimposed information, which might be an approach to fusing collections as mentioned in item 6.
10. How do addressing modes and related capabilities (address comparison, value-to-address mapping) in the base space affect the ability to structure and process superimposed information? Or turning the question around, how can base information be structured and managed to better support superimposed information?
11. Some current information spaces currently have quite rich addressing mechanisms, such as URLs and XPointers for the web. Modes of external addressing for other kinds of information aren't as well developed. For example, what are reasonable ways to identify information elements in relational or object-oriented databases? Is it possible to support addressing in virtual data, such as web pages generated on the fly from data stored in a different form?

12. How can existing database management systems be extended to be able to interact with data they don't store, and thus serve as a better basis for superimposed information management?
13. By what means can superimposed information spaces be populated automatically, or with minimal human effort? What are possible mechanisms for maintaining superimposed information in the face of changes in the base level?
14. What are good formats for the representation and interchange of superimposed information? XML seems an obvious choice to investigate. Are there any other likely candidates?

7 Acknowledgements

We would like to thank Peter Buneman, Rick Hull, Serge Abiteboul and Sophie Cluet for discussions on this topic. This research is supported in part by NSF grant IRI 94-98191.

References

- [BP98] Brin, S. and Page, L., "The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Proceedings of the 7th World-Wide Web Conference*, 1998.
- [DM99] Delcambre, L. and Maier, D., "Models for Superimposed Information," unpublished manuscript, May 1999.
- [DMRA97] Delcambre, L. Maier, D., Reddy, R., and Anderson, L., "Structured Maps: Modeling Explicit Semantics over a Universe of Information", L., *Journal of Digital Libraries*, Vol. 1, No. 1, 1997.
- [RDF99] Resource Description Framework (RDF) Model and Syntax Specification, W3C Recommendation 22 February 1999, <http://www.w3.org/TR/REC-rdf-syntax/>
- [TNM99] ISO/IEC 13250, "Topic Navigation Maps", <http://www.infoloom.com/topmap.htm>
- [XL98] XML Linking Language (XLink), World Wide Web Consortium Working Draft 3-March-1998, <http://www.w3.org/TR/WD-xlink>
- [XML98] Extensible Markup Language (XML) 1.0, W3C Recommendation 10-February-1998, <http://www.w3.org/TR/REC-xml>
- [XP98] XML Pointer Language (XPointer), World Wide Web Consortium Working Draft 03-March-1998, <http://www.w3.org/TR/WD-xptr>