[DigLib] Stanford Digital Library Project
Working Paper

Title: Modes of Information Integration
Author: Terry Winograd
Abstract: In order to better understand the different approaches to digital libraries, I compared a number of existing and proposed systems and developed a taxonomy that can be used in identifyingthe different tradeoffs they make in the overall design space.

Contents:

Modes of Information Integration

In order to better understand the different approaches to digital libraries, I compared a number of existing and proposed systems and developed a taxonomy that can be used in identifyingthe different tradeoffs they make in the overall design space.

1. FULLY INTEGRATED DOCUMENT REPRESENTATION

In this class of systems, the entire system with all of the information is managed by a uniform suite of programs. Typically both the document formats and the meta-information structure is specific to the system design. The early hypertext systems such as Xanadu and NLS (later Augment) were of this kind. Later commercial products such as Xerox STAR, Lotus Notes,

These systems grew up along with a utopian vision of everyone (or at least everyone in the enterprise) buying into the system and putting their information into its form. This made it possible to provide highly integrated interfaces and services.

A more recent class of systems like this are intended for more limited use. Rather than handling all information needs, they focus on providing a structured body of informationin a uniform way. Systems such as Bellcore Superbook, SUN Answerbook, MIT TechInfo, and many SGML-based systems fit this model. They started out from large collections of documents within a company for a particular purpose (e.g, System documentation, policy book, ...). Documents are created in special form (or converted with structural information included in the conversion). Information services (e.g., search and bookmarks) are integrated into the system architecture and interface. These systems can provide well-integrated facilities for acces control, search, bookmarks, history, checkin/out, etc.

A fuly integrated system can be distributed across multiple servers (as is

Lotus Notes and some of the documentation systems). The unification is in the materials and the coordinated management of those servers.

2. INTEGRATED FILE SYSTEM

As the microcomputer industry proliferated the number of different formats and applications, it became less viable to assume that the relevant information for a user would all be in one uniform well-controlled structure. What was needed included a way to coherently manage a set of files in multiple formats and across multiple servers. The coherence was in the way that files were stored, cataloged in directories, copied, searched, format-converted, etc.. An ordinary file system does much of this locally, and there are a number of systems that do it on a distributed basis (e.g., the DEN architecutre of Xerox and Novelll and the Andrew File System). There is a centralized object/file storage system and individual information resources can be in native format. To have things in the system, they must be entered into it in the same sense as putting a file into a file system. A specific set of services are provided by the file-system servers, including format conversion, access control, source and version control, etc.

3. INTEGRATED INDEX

As it became more commont for an individual to access materials located at multiple sites on the Internet, there was increasing need to be able to find things without haeing a specific file system in which to look. A number of projects have developed indices that store various kinds of material (file names, descriptors, word frequency indices,...) that enables search across multiple sites. These include WebCrawler, Archie, Veronica, and others that require no special cooperation on the part of the site, and others, such as Harvest which can get more specialized information (meta-information about documents) by having sites provide appropriate descriptions. Some grew out of the world of information retrieval, with its common indices and search over collections of sources. Others, such as Yahoo, provide a structured cataloging service of distributed materials.

4. INTEGRATED SERVICE CENTER

The advantagesof integrated file systems is that they can provide a base for serices that go beyond simple search. The disadvantage is that they only deal with materials stored in the system. A mode of integration that combines the advantage of integrated file systems with those of distributed materials is what I will call the integrated "service center." In these systems (e.g., InfoHarness, GAIA, Interspace, ...) There are centralized meta-information databases (usually object-oriented). That contain structured information about sources that may be on the same server or elsewhere. One motivation for these systems came from the desire for "enterprise-wide" resource rationalization, in which people in a company want to have uniform access to information in a variety of storage forms and servers, but don't want to (or can't) enforce an integrated document representation.

Resources can be anywhere, cataloguing is done explicitly, in forming appropriate "collections" and extracting (automatically or semi-automatically) meta-information about the objects in them. Typically there is an "administrator" who is responsible for putting the picture of the distributed sources together. Entering of materials into the catalog is an activity independent of creating and storing the materials. Access to the sources by users may be via the integrated system, or may be shortcut by having it give out direct access addresses to the end-user interface client.

There is no limit to the kinds of meta-information or services that can be provided, such as search (of different kinds), source and version control, access management, translations, archiving, backup, etc.

5. INTEGRATED NAME SPACE

A simple form of integration that has proved extremely powerful in the distributed Internet is the use of a uniform space that allows directly usable identifiers for information objects anywhere on the net. The combination of the Domain Name System (for hosts) and the Uniform Resource Locator (for objects on that hosts) has made possible the explosive growth of the WorldWideWeb. At this level of integration, all that is managed is the name space, not the contents of objects being named.

6. INTEGRATED SERVICE PROTOCOLS

The final kind of system is the most loosely coupled, in that both sources and services are distributed, using a common meta-language and protocol for integration. This is the architecture for two of the DLI projects, at Michigan (with its "conspectus") and Stanford (the "infobus"). In these systems, independent sources and sercices are described using a common language and ontology, alllowing the indpendent parts to communicate with one another to find, apply and present serices of a variety of kinds.

These are different from the integrated service centers in that they do not use a centralized body of meta-information managed by an aminsitrator who controls what is "entered" into the system, but instead attempt to make use of istributed metainformation provided by a combination of parties, including the originators of information, the service providers who use it, and third parties.

The architecture does require a centralized body of "model" information for use by distributed services. This defintes the structure of the meta-information for sources and service activities. It could either be declared once and for all, or (as in the current projects) is extensible and can be changed over time.

The following table summarizes the above and makes some claims about the current DLI projects. Numbers indicate notes that folllow:

System	Integr. type	Doc. format	Store mgmt.	Meta storage	Meta format
Stanf.	Protcol	no	no	no	yes
Bkly.	Center	some	some	yes	yes
Mich.	Protocl	no	no	yes	yes
CMU	Full[1]	yes	yes	yes	yes
Sta.B.	Full[2]	yes	yes	yes	yes

Ill.	Center	no	no	yes?	yes

1. CMU is a fully integrated system for video and correlated data. It may make use of other information forms as well.

2. Santa Barbara is a fully integrated system for geographic data. It may make use of other information forms as well.

Change history:
HTML version June 24, 1995 by Terry Winograd
First draft April 25, 1995 by Terry Winograd