NCSTRL: Networked Computer Science Technical Reports Library

NCSTRL (pronounced "ancestral", which I think is supposed to connote a bountiful and distinguished progeny) is a collection of computer science technical reports from CS departments, made available for non-commercial and educational use. It is an outgrowth of two previous projects in digital libraries.

1. DIENST at Cornell (funded by ARPA)

   NCSTRL uses the Dienst software, an open, distributed, implementation of digital library servers. Dienst provides three basic services at each site:

   - a repository
   - indexes and a search engine
   - a user interface.

   The repository can hold documents in different formats (PostScript, TIFF, GIF, HTML) and can handle the same document in multiple formats. Sites running Dienst can use home-grown software for any of these components. For example, they can supply their own browsers (and some do).

   Dienst makes a distributed library appear as a single collection at the top level, even though the repositories, search engines, and browsing interfaces are distributed.

2. WATERS at Virginia (funded by NSF)

   WATERS is (was) a distributed database of computer science technical reports kept at different sites, with a centralized index of bibliographic information (including abstracts) in REFER format that is supplied by librarians at each site.

In NCSTRL, participating organizations choose either to run the DIENST software ("Standard" NCSTRL), or merely maintain an FTP repository and contribute bibliographic information (WATERS-fashion) to a special DIENST server ("Lite" NCSTRL). The Lite connection is easiest but the Standard connection allows a site to maintain its own search engines, to provide access to documents in multiple formats, and to provide customized user interfaces -- thumbnail browsing, page zooming, page selection, support for logical document structure, etc. NCSTRL currently (July 1996) has 55 participants,

- 20 Standard sites
- 35 Lite sites, served through the University of Virginia.

To access the collection you can either examine a specific participating site's reports, or do a simple search, or perform a "fielded" search. Sites list technical reports by author and by year (although Standard sites can format the lists however they want). A *simple* search is an OR query on author, title, and abstract fields. A *fielded* search allows you to specify terms for one or more fields and decide whether the fields (not the terms) should be ANDed or ORed, and select a subset of sites to be searched. Terms--even author names, according to the documentation--are stemmed.

Comments

NCSTRL has a very different philosophy to the New Zealand Digital Library.

- Sites participate, either by supplying bibliographic information about their FTP repositories, or by running their own indexes, search engines, and browsing interfaces.

  This has the advantage that bibliographic information is likely to be accurate, but the drawback that it is necessary for institutions to subscribe to the service and devote resources to it. Will institutions continue to do this indefinitely? Lack of uniformity creates a problem: some sites do not supply abstracts; sites can choose to keep full reports unavailable.

- The architecture is "scalable" in principle because of the parallelism achieved by several sites running Dienst.

  Scalability is a clear advantage, but the architecture encourages lack of uniformity--some sites have quite different interfaces. Will sites continue to run the Dienst software in the long term?--on a random day the software proved to be down on three of the twenty sites.

- Indexing is like a standard bibliographic library catalogue supplemented by the possibility of matching keywords in titles and abstracts, except that the physical sites play a more prominent role in searching and browsing than in traditional libraries.

  The lack of a full-text index is a significant disadvantage when searching the collection, but even more serious is the poverty of the query language. I found it impossible to fulfil my standard request for documents about *optical music recognition* because the implied OR returned several hundred documents and there was no way of refining the query.

Ian H. Witten
July 1996