## World-Wide Web: What Links to databases?

Jacques Guyot[[1]](http://cui.unige.ch/db-research/Enseignement/analyseinfo/WWWSOUG1E/WWWSOUG1E_fn.html#fn0)

Ian Prince[[2]](http://cui.unige.ch/db-research/Enseignement/analyseinfo/WWWSOUG1E/WWWSOUG1E_fn.html#fn0)

# Introduction

FTP, TCP/IP, WWW, HTML, HTTP: these acronyms have all received large press coverage recently. Terms such as the Internet Highway are rapidly becoming household names. Useful or gadgets? Stable or transitional technologies? What new domains are emerging? What relevance to databases? To answer, at least partially, some of these questions, we will be concentrating our discussion on the World-wide Web (WWW) and its relationships to databases. The discussion is divided into three parts:

1. In the first part we will discuss the basic concepts of networks and hypertexts and the necessary hardware and software. We will describe a compete example of hypertext generation for the SQL and Oracle's PL/SQL languages.
2. In the second part we will describe the concept of HTML forms for WWW data acquisition and queries. We will demonstrate how forms allow interacting with an Oracle database via the WWW.
3. In the third part we will take an opposite approach and consider how hypertexts can be generated from databases. Again an example will illustrate the basic concepts.

Parts II and III will be published in subsequent issues.

# 1. Networks and hypertexts

The World-Wide Web combines two concepts: the concept of an interconnected network (the Internet) with the concept of hypertext. Both concepts have separate origins and purposes. Their combination opens up new areas that until recently were unthinkable.

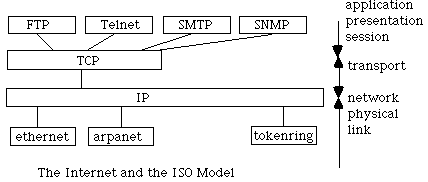
We will first define a few key concepts.

## Interconnected Communication Networks (the Internet)

The Internet's origins can be traced back to 1968 and the ARPANet (Defense Advanced Research Project Agency Network). The concept of interconnected networks allows running different protocols on different networks, requiring an interconnection protocol that runs above each network protocol. IP (the Internet Protocol) defines the unique addresses for the network and its hosts. The IP protocol fulfills two functions:

1. packet routing through the different networks. A routing can send a packet to either the receiver or to the next network gateway;
2. the fragmentation and assembling of packets that are too long.

The network itself assures packet transport. IP packets are encapsulated by the sender and each time a packet leaves an IP gateway. Likewise, IP packets are decapsulated by the receiver and each time a packet enters an IP gateway. The IP protocol is relatively simple and robust and has certainly contributed to the Internet's success.



**TCP** (Transmission Control Protocol) is the transport layer above IP that ensures the connection between sender and receiver. TCP provides a virtual circuit and error detection/correction. Using TCP/IP two network hosts can communicate independently of the networks that connect them.

Above TCP/IP application oriented protocols provide higher level services.

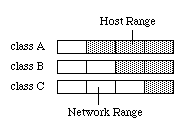
**FTP** (File Transfer Protocol) allows file transfers between two network hosts. FTP servers such as nic.switch.ch provide vast collections of documents, programs and images. To overcome long-distance FTP traffic many popular FTP servers are geographically duplicated (nic.switch.ch "mirrors" many of the most popular US FTP servers). A Swiss user of the Internet will most certainly find better FTP performance connecting to a Swiss "mirror" than directly connecting U.S. server. ARCHIE servers are essentially databases that index documents from different FTP servers, largely facilitating finding where a certain file is located.

**Telnet** (Terminal Protocol) is the terminal access protocol. Using Telnet it is possible to run remote programs (such as databases) using a command-line interface.

**NNTP** (Network News Transfer Protocol) is the bulletin board protocol. NNTP allows an Internet host to participate in any of the thousand of forums ranging from technical discussions on Oracle databases (comp.databases.oracle) to cultural discussions on Switzerland (soc.culture.switzerland).

**SMTP** (Simple Mail Transfer Protocol) is the protocol permitting electronic mail transfers between Internet hosts.

**SNMP** (Simple Network Management Protocol) is the protocol to manage networks. The SNMP protocol is usually built in to network hardware such as gateways, routers and multiplexers.

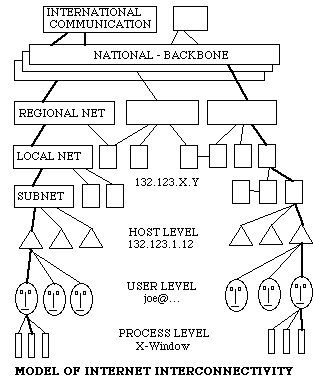


IP addresses consist of 4 octets whose values are separated by dots, for example: 136.102.233.49. The first octets provide network addressing and the last octets the host addressing. The balance between network and host addressing is determined by the class of network: A, B, or C.

The InterNIC Registrar is responsible for the distribution of network addresses. The hierarchy is usually the following:

1. - geographical zone
2. - organization
3. - department
4. - host

Name servers allow communicating logical addresses rather than physical addresses. For example, cui.unige.ch is the logical address of the CUI's principle host. To determine an electronic mail's address the name of the person is concatenated to the host's, for example guyot@cui.unige.ch or prince@acm.org



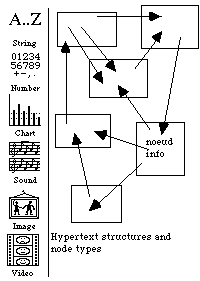
The number of Internet users can only be estimated. Currently (March 1995) it is estimated that there are almost 5 million hosts connected to the Internet with an annual growth rate of 50-100% (approx. 10'000 additional hosts per day). More than 10 Terabytes are currently exchanged monthly on the U.S. backbone of the Internet.

## Hypertexts

The idea of associative links can be traced back to Vannevar Bush and Douglas Engelbart.

The associative link consists of a destination node and a source node, where both nodes usually reside in different documents. Using links it is possible to create a semantic web between information nodes. A reader of such documents can follow links depending on his or her interests.

Systems such a HyperCard and Oracle's on-line help already implement a simple form of hypertext, where the user navigates from one explanation to the next. More recently CD-ROMs use the same concept of hypertext navigation to present large quantities of multimedia (text, sound, images) information.



## The World-Wide Web

The WWW is a hypertext where the nodes are distributed throughout the Internet. Nodes can contain text, images, sound, or animations. The WWW, therefore, combines two webs: one physical (the Internet) and one semantic (the interconnected documents).

WWW browsers provide the user interface to view and navigate the hypertext. Links within a node are distinguished typographically (by underlining, for example). Clicking on a link presents the document that the link "points" to.

WWW documents, essentially, do not define how a document is to be viewed. This has allowed the development of browsers on different platforms (Windows, MacOS, UNIX, VMS) such as NCSA Mosaic, Netscape and Lynx. Each browser determines the presentation of the structural information of a WWW document.

### HTML

Hypertext documents on the WWW are written in HTML (Hypertext Markup Language), a relative of SGML (Standard Generalized Markup Language).

The following shows an example of HTML use.

<TITLE>Jump in the Web</TITLE>

<H1>Once Upon a Time ...</H1>

Welcome to the world of HTML

this is the first paragraph.<P>

this is the <B>second</B> and <I>some URL</I><P>

<A HREF="http://oracle.com"> at Oracle </A><P>

<A HREF="http://cuiwww.unige.ch"> at SQL7 </A><P>

<A HREF="http://mistral.nst.fr/~pioch/louvre"> Le Louvre </A><P>

<IMG SRC="smile.gif">

Comments:

Delimiters are specifies by angled-brackets <like this> and generally go in pairs: <xxx> et </xxx>

**<TITLE>**Jump in the Web**</TITLE>** Defines the title of a document node

**<H1>**Once Upon a Time ...**</H1>** Defines a header of level 1 (out of six)

Welcome to the world of HTML this is the first paragraph.**<P>** Defines a paragraph

this is the **<B>**second**</B>** and **<I>**some URL**</I>**<P> Defines text ornamentation (bold and italics)

**<A HREF="http://oracle.com">** at Oracle **</A>**<P> Defines a link to the Oracle HTTP server.

**<IMG SRC="smile.gif">** Adds an image to the document

The following figure shows how the page is rendered by NCSA Mosaic.

### WWWSOUG1E5 URL

A URL (Uniform Resource Locator) locates a unique WWW document on the Internet. It identifies the type (HTTP, FTP, GOPHER, etc.), the host, and the path on the host. A URL takes the following form:

type://host:[port]/path#anchor

Anchors define a specific point in a HTML document. In Part II of this article we will show how URLs can be used to access databases via the CGI (Common Gateway Interface) mechanism.

### HTTP

A HTTP server (Hypertext Transfer Protocol) is responsible for serving documents (specified by URLs) to clients (browsers). HTTP implementations exist on most platforms (PC, MacOS, UNIX, etc.). HTTP servers are relatively simple and modestly priced (if not totally free).

## Lost in hyperspace

In 30 months, HTTP traffic has increased from 10Mb/month to 1Tbyte/month, equaling FTP traffic. In the last 12 months the number of HTTP servers has increased tenfold currently (March 95) reaching. 10'000. Millions of WWW browsers have been downloaded. Most Fortune 500 companies have, or are developing, their own servers (maximum visibility for a minimum cost!). This abundance has its side-effects: "Everything is out here on the WWW, but how can I find it?"

There are basically two methods for finding information on the WWW: (1) by using compiled indexes such as the CUI's W3Catalog or (2) by serendipitously navigating links also know as "net surfing".

## Generating hypertexts

HTML hypertexts may be generated by simple editors, by specialized HTML editors, or by structured document converters (RTF to HTML, for example).

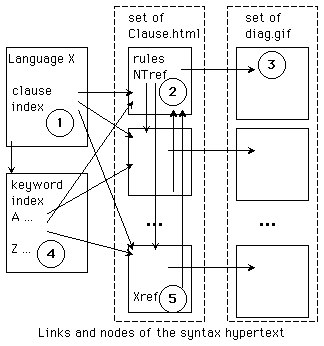
Here we will demonstrate a specific case of structured document conversion, where the document to be converted represents a language grammar.

### Syntax hypertexts

Our goal is to provide syntactical guides to programming languages such as SQL and Ada to students and developers.

BNF (Backus Naur Form) grammars of these languages are commonly available. As such we have developed tools that take a BNF grammar as input and produce an HTML hypertext of the syntax as output.

The syntactical structure of the language is defined by clauses (1). Each clause generates a textual hypertext node (2) and a corresponding syntactic diagram (3). Each clause is described by a number of keywords and reference to other clauses (the non-terminals of a clause). Each non-terminal generates a hypertext link. A keyword index is also created (4) with links to each clause it appears in. Finally all the references made to a clause are generated (5).



Take for example the following PL/SQL clause:

plsql\_block=

["<<" label\_name ">>"]

["declare" declare\_spec <declare\_spec >]

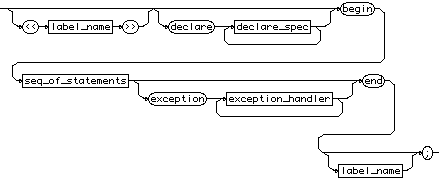
"begin"

seq\_of\_statements

["exception"

exception\_handler <exception\_handler>]

"end" [label\_name] ";"

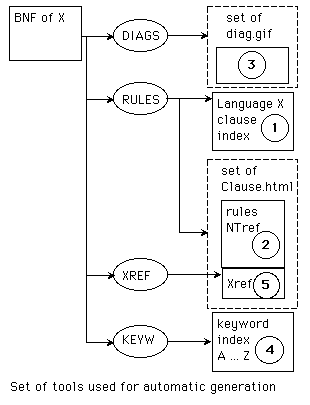


The following is required: each non-terminal (label\_name, declare\_spec, seq\_of\_statements, exception\_handler) links to its corresponding clause; that the keyword "end" refers to plsql\_block as well as function\_body, if\_statement, etc.; that this clause is referred to by (among others) the create\_trigger clause. D

For a language such as SQL for Oracle7, one must count on approx. 250 clauses, representing many thousands of links. Automatically generating the hypertext eliminates possible errors and also allows re-generating the hypertext when the grammar changes.

We have developed a family of tools for such hypertext generation from BNF grammars:

**DIAGS**: draws the syntactic diagrams given a BNF grammar. The BNF grammar is actually converted into drawing primitives that are then interpreted to produce the diagram itself.



**RULES**: creates both the clause index and the nodes for each index. Each keyword is displayed in bold and each non-terminal creates a link. The corresponding syntactic diagram is also added to the clause-node.

**XREF**: creates an index of all the calling clauses and adds these to the clauses creates by RULES. This technique can be extended to other information pertaining to clauses, such as explanations, examples, etc.,.

**KEYW**: creates an index of keywords and defines all the links to the corresponding clauses.

## WWWSOUG1E9

SQL grammar viewed with NCSA Mosaic.

We have currently used the toolset for generating hypertexts for SQL, PL/SQL et ADA[[3]](http://cui.unige.ch/db-research/Enseignement/analyseinfo/WWWSOUG1E/WWWSOUG1E_fn.html" \l "fn0).

The most difficult part of this technique is describing the languages' BNF grammars. The decomposition of the grammar into clauses depends much on the goals of the hypertext. For the Ada language, the decomposition is the same as that of the Ada standard. For SQL and PL/SQL we have taken a different approach, trying to regroup a language concept in each clause (although the decomposition still resembles the one chosen by Oracle...). For a language the size of PL/SQL, one must count on two days work in defining the grammar and half a day to generate the hypertext.

## Conclusions

We believe that developing WWW servers, whether private public, can greatly benefit and reach a large number of users.

WWW Servers can reflect a companies image. A number of applications are still to be discovered, but one can already cite:

1. company "Who's Who" directories;
2. courseware and tutorials;
3. internal procedures and standards;
4. collaborative agendas.

The principle advantages of the WWW is its user-interface simplicity. With one tool one can access information from multiple sources, independent of location, software or hardware.

For the first time a unified vision of information is possible.

## Relevant URLS

http://cuiwww.unige.ch/db-research/Enseignement/analyseinfo/index.html

access point to the BNF hypertexts for SQL, PL/SQL, and Ada.

http://cuiwww.unige.ch/db-research/Members/jg/jg.html

http://cuiwww.unige.ch/~prince

information about the authors

http://cuiwww.unige.ch/

the CUI server at the Univerity of Geneva

http://cuiwww.unige.ch/cgi-bin/w3catalog

the CUI `s main index to the WWW

## Bibliography

Communication of the ACM, Vol. 33, No. 3 Special Issue on Hypertext (1990)

Communication of the ACM, Vol. 37, No. 2 Special Issue on Hypermedia (1994)

Communication of the ACM, Vol. 37, No. 8 Special Issue on Internet Technology (1994)

The World-Wide Web, J. December and N. Randal, Sums Publishging (1994)

Mastering the Internet, G.H. Cady and P. McGregor, Sybex (1995)

E. Krol and M. Loukides, The whole internet: user's guide and catalog, O'Reilly (1992)