

Accessing HEP Collaboration Documents using WWW and WAIS

Trang D. Nguyen, Elizabeth Buckley-Geer, David J. Ritchie

*Fermilab, P.O. Box 500
Batavia, IL 60510, USA*

We describe a system for accessing HEP collaboration documents using WWW and WAIS.

1 Brief History of WAIS and WWW

WAIS stands for Wide Area Information Server¹. It is a distributed information retrieval system. A WAIS system has a client-server architecture which consists of clients talking to a server via a TCP/IP network using the ANSI standard Z39.50 V1 protocol. WAIS was originally developed by a team at Thinking Machines, Inc. The leader of the project, Brewster Kahle, later formed WAIS Inc. which provides commercial WAIS software and services. A freely available version (FreeWAIS) is supported by the Clearinghouse for Networked Information Discovery and Retrieval, also known as CNIDR².

FreeWAIS-sf, which is the software we are using at Fermilab, is an extension of FreeWAIS. It is being maintained by Ulrich Pfeiffer at the University of Dortmund, Germany. FreeWAIS-sf supports all the functionalities which FreeWAIS offers as well as additional indexing and searching capabilities for structured fields.

World Wide Web (WWW) was originally developed by Tim Berners-Lee at CERN³ and is now the backbone for serving information on Internet.

2 The Collaboration Document System

With the explosion of the web world wide, we recognized the need to take advantage of this medium to share documents among collaborators on high energy physics experiments. Combining FreeWAIS-sf and the CERN httpd server, a mechanism was put in place to search collaboration documents using catalog information such as title, author, date and/or to perform string searches against the content of these documents. Implementations of this system have been put in place for the CDF and D0 collaborations.

3 What are the major components of FreeWAIS-sf and WWW?

In order to create a searchable FreeWAIS-sf database on the web, it is important to identify the following items:

- a collection of documents to index
- a catalog of the document collection. (See Figure 1)
- a web server to host the WWW/WAIS activities

FreeWAIS-sf supports a variety of file formats. These files can be stored anywhere on the network so long as they are visible to the operating system where WAIS is installed. In the case of the WWW/WAIS implementation at CDF, the files are stored on a VMS cluster in Postscript or plain text format. The VMS disk is NFS mounted on a Unix machine where the web server is located and FreeWAIS-sf is accessible.

At CDF, the majority of documents are written in TeX and many contain embedded graphics. Although WAIS supports the indexing of a variety of file formats such as sound, pictures, and video, we only utilize its postscript and text indexing capabilities.

At CDF, the catalog for the documents is stored in a Datatrieve database on the VMS cluster. This is the existing system established some years ago by which documents were (and continue to be) tracked. This provides the document numbers. A listing of the database is extracted and saved in a file which is then indexed by FreeWAIS-sf's utility, waisindex.

```
NU: 5
AU: Grannis, P.
TI: Effect on Z-Zero Width from Overlapping Hits
DT: 07/31/83
FL: /d0sg10/data0/WWW/docs/physics_analysis/d0notes/source/note5.ps
AL
NU: 6
AU: Grannis, P.
TI: Slides Shown at the Preview of D0 Review
DT: 06/01/83
FL: /d0sg10/data0/WWW/docs/physics_analysis/d0notes/source/note6.ps
AL
NU: 7
AU: Grannis, P.
TI: Side View E740 End Cap Electromagnetic Detector
DT: 8/29/83
FL: /d0sg10/data0/WWW/docs/physics_analysis/d0notes/source/note7.ps
AL
NU: 8
AU: Grannis, P.
TI: Comparison of Hadron Rejection Efficiency
DT: 9/4/83
FL: /d0sg10/data0/WWW/docs/physics_analysis/d0notes/source/note8.ps
AL
NU: 9
AU: D0
```

Figure 1: Example of a structured field entry.

To keep it simple, we created two WAIS databases. One WAIS database is used for content searches and the other WAIS database is used for catalog searches. The primary key which is used to link databases together is the document number. Users post their documents to the VMS disk using an automatic procedure. This ensures that the document number is always part of the filename and allows traditional file and directory searches to be used when needed (e.g. for scripted searches and processing.)

Under the existing system, there is no link between Postscript files and the catalog entry in Datatrieve. Under WWW/WAIS system, the WAIS search utility, waisq, is invoked in a perl CGI script. In this way, a user invokes a web form to pass queries (Figure 2) to search the WAIS database and return catalog information as well as hypertext links to the Postscript files (Figure 3). Older catalog entries do

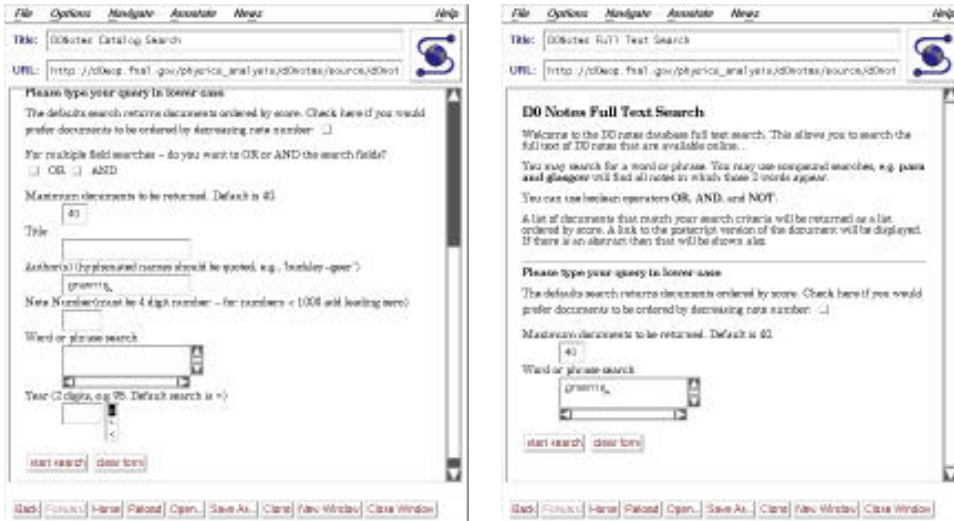


Figure 2: Examples of a catalog and a full text search interface.

not have Postscript versions of the document, in this case just the catalog entry is returned.

At D0, the implementation is similar. D0 documents are stored in Postscript on VMS and are accessible on a Unix system via NFS. A FileMaker Pro database on a Macintosh is used to maintain and assign new document numbers.

4 Usage at Fermilab

The CDF and D0 collaborations at Fermilab are making significant use of this system to search their internal document database. The CDF collection consists of publication materials, descriptions of detector implementations, various stages of data analysis, etc. The average number of searches of the WWW/WAIS database is 300 per month.

The D0 collection similarly consists of transparencies for D0 meetings, technical notes, draft of publications, internal design documents, manuals, etc. The WWW/WAIS database is still new to the D0 collaboration. It is too early to determine the average number of searches. However, we have received positive feedback from many D0 users.

Based on our accounting data, the CDF collaboration perform more catalog searches than content searches using the WWW/WAIS system. Prior to the WWW/WAIS technology, the mechanism to perform catalog searches for the CDF document collection existed; however, it was accessible only on the VMS cluster where the Datatrieve database was installed.

At D0, however, the feedback we have received suggests that the content search capability of WAIS is more popular. Without the WWW/WAIS system, catalog searching for the D0 document collection is primitive.

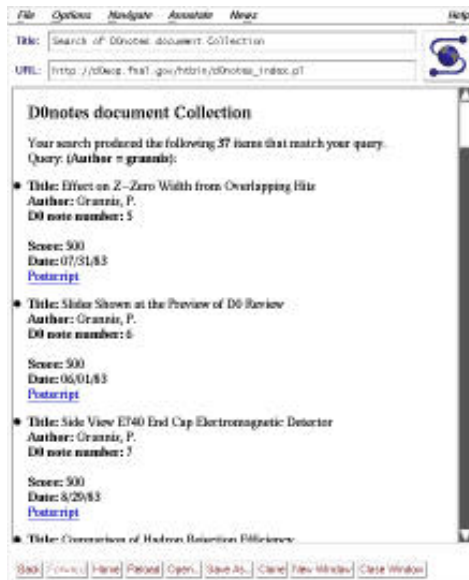


Figure 3: Example of hits returned.

Based on anecdotal comments, we believe that CDF and D0 users are finding this mechanism useful in supporting their research. Work is in progress to create a mirror site in the CDF case to support CDF collaborators in Taiwan. We plan to make a WWW/WAIS kit available via the Unix Product Distribution (UPD) at Fermilab to benefit other collaborations.

5 What are the advantages?

The cost is right (i.e. free). The WWW server and FreeWAIS-sf are freely available on many platforms. It is in keeping with the tradition in HEP to search and make use of any useful tools which are available in the public domain. Good examples for this are TeX, emacs, and now www.

It is accessible. Since the web is used, the collaboration document system is accessible worldwide.

It handles internal documents. It is highly desirable to be able to share internal notes within the collaboration. Security control has been implemented using the http security mechanism bundled with the web server. At CDF, the documents are password protected; at D0 the documents are hostname (ip) protected from the web.

FreeWAIS-sf is versatile. It can index a wide variety of file formats which include text, postscript, dvi, html, emacsinfo and more. It works with many image formats such as gif, pict and tiff by indexing filenames. Compared to the functionality of commercial vendor products (e.g., Topic from Verity, Inc.), FreeWAIS-sf indexes postscript files with ease.

The user interface is platform-independent and non-proprietary. The web form can be handled by many web clients which support forms. The designer has total control over the interface, up to the limitations of the HTML language. In addition, it is transparent to users where the files are stored allowing flexible management of document files.

Finally, searching a WAIS database can be invoked at the command line. Therefore, it can be called from a CGI script using perl, shell language, C, or other high level programming language.

6 What components are lacking?

Major components which this system lacks are the ability to assign document numbers, enter catalog information and track document workflow and signoff. A separate database is still needed for the document number assignment. Similarly, local techniques have to be developed to enter the catalog information and track document workflow and signoff.

However, this is not necessarily a disadvantage. The ways for assigning numbers, collecting catalog information, etc., vary widely from collaboration to collaboration while, with the advent of the web, the methods for disseminating information about document catalogs and document content are now quite uniform. Collaborations may tailor the system to their unique local procedures while providing wide access to their documents in a uniform way.

7 Conclusion

The combined implementation of WWW and WAIS for accessing high energy physics collaboration documents has resulted in a useful tool to make collaboration documents accessible in a cost-effective and convenient manner.

8 Acknowledgements

We would like to acknowledge the helpful advice and comments of Eric Wicklund for bringing WAIS to our attention. We would also like to acknowledge all the work and efforts of Adam Para for helping to bring the WWW/WAIS implementation to the D0 collaboration. Lastly, we are very grateful for advice and assistance from Ulrich Pfeiffer, the author of FreeWAIS-sf.

References

1. Pfeiffer et al., "FreeWAIS-sf," Sept 9, 1994
2. Pfeiffer, "WAIS-FAQ," 1.2 1994
3. Cailliau et al., "The Use of the World Wide Web in HEP", Proc. Conf. on Computing in High Energy Physics '94, Berkeley, CA, April 1994