COADS Metalog: Dynamic COADS Documentation

William R. Moninger
Forecast Systems Laboratory

Scott D. Woodruff
Climate Monitoring and Diagnostics Laboratory

Ralph J. Slutz
Independent Contractor

Richard L. Fozzard
Cooperative Institute for Research in
Environmental Sciences

Abstract

Modem computer technology offers text management techniques that can be helpful in documenting environmental science data sets. We have applied some of these techniques to the Comprehensive Ocean- Atmosphere Data Set (COADS), a large data set widely used in global change research. We have built a system called COADS Metalog that makes a large amount of COADS documentation (metadata) available in electronic form, and a system called Metalog that can be used to document any data set. We describe Metalog, COADS Metalog, as the process by which we built the systems, and discuss lessons learned.

1. Introduction

We have begun to apply new computer technology to the management of environmental metadata (information about data). As an initial application, we chose the Comprehensive Ocean-Atmosphere Data Set (COADS) because it is important to global change research, is widely used, and is generally well-documented. Also, it is less homogeneous than many environmental data sets, having variable data density in time and space, variable data gathering procedures, and varying levels of data certainty; thus it is a particularly challenging data set to address, and one for which the benefits of computer documentation management are likely to be substantial.

1.1 Problems with metadata

Until recently, environmental data sets were used primarily by scientists who were familiar with how the data were gathered and processed. Because of this familiarity, thorough metadata was not needed. Now, however, because of growing social concern with environmental problems and the increasing expense of gathering new data, the environmental science community is being encouraged to assemble data sets that can be widely used. Users who have immediate access to those who assembled the data can receive the information they need to avoid pitfalls through

personal contact. Other potential users are not so lucky; for them, metadata may be deficient in several ways.

- Important metadata may not be readily available. They may be in handwritten logbooks stored in the offices of individual researchers, or in informal technical reports circulating among a limited readership. Such metadata are difficult to identify, duplicate, and acquire.
- Metadata that reach the refereed literature, though authoritative, may lack important detail, and may be out of date.
- When a large amount of metadata are available, not all of them will be relevant to specific research tasks. A new data user may miss critical items of information that are buried in irrelevant detail.
- Metadata are often written using a technical vocabulary that may be unfamiliar to scientists in different fields than the author. This problem is increasing as investigators with wide-ranging areas of specialization participate in global change research.

1.2 COADS Metadata

A wide variety of metadata exists for COADS, very little of which is now readily available in digital form or from any single archival source. These metadata can be roughly divided into two general categories. The first of these is "textual and graphical" information. Within this category a further distinction is useful according to the formality of the information and the level of peer review to which it has been subjected.

Such information may range from the least formal (e.g., "back of the envelope" notes, electronic mail or bulletin board messages, etc.), to internal or selectively distributed publications (e.g., technical reports), up to the most formal (refereed literature, books). The less formal and more free-form elements of this information category are commonly referred to as "gray literature"; the more formal and less free-form elements are likely themselves to contain overall structures such as a carefully defined logical progression (sections, chapters, etc.), tables, figures, etc.

The second major category of metadata is much closer in some sense to "real data." We will refer to these as "structured metadata," although the textual/graphical category also frequently would have some intrinsic structure. The elements of this second general information category may be more likely to be readily available in digital form, and may possess a structure that can be completely represented, for example, in tabular form. Due to this homogeneous structure, such metadata may be efficiently stored, updated, and retrieved using available database technology.

One example of structured metadata for COADS is the annual tables of individual ship instrumentation details and other characteristics available from WMO publication 47 (1955-), currently available in digital form since about 1973. Using these metadata in conjunction with individual ship reports, a user may be able, for example, to adjust winds based on an individual ship anemometer height or to better separate sea surface temperatures taken by bucket from those taken by engine room intake. Other examples of structured metadata for COADS include similar instrumentation details that may become available for buoys (see Woodruff et al., 1991 for a discussion of the availability of such buoy metadata), or an address list of other COADS users that

might allow users to interact with others who have obtained similar COADS products or who share similar research interests.

1.3 Advantages of electronic metadata

Metadata that are in electronic form have several advantages over the traditional printed form. First, electronic metadata can be copied and distributed easily, either by mailing floppy disks, or by e-mail. Second, electronic metadata can be rapidly searched for particular words and phrases. Third, electronic metadata can be cross-referenced easily and extensively. So even when two items of metadata do not share a similar vocabulary, and hence might not be retrieved together in a simple

word search, they can be retrieved via electronic "links" that have previously been established. Fourth, because distributing, searching, and cross-referencing of electronic information is so inexpensive, it is feasible to share gray literature. Fifth, and most important, when metadata are in electronic form, advantage can be taken of new information management techniques recently developed by the large and active computer science community.

1.4 Relevant activity in computer science

Although most environmental scientists think of computers as numeric processors, many computer scientists view computers primarily as symbol manipulators. Printed words are symbols, so it is not surprising that computers can successfully manage text. Word processors are familiar examples of text-management software, but far more sophisticated systems are now in use. Some examples of state-of-the-art text management systems are the following.

> CONSTRUE (Hayes and Weinstein, 1990) is a system used by Reuters News Service to automatically classify all their news stories into one or more of 674 categories. CONSTRUE is faster and more accurate than human indexers.
>
> SCISOR (Jacobs and Rau, 1990) automatically identifies and summarizes documents in particular subject areas. SCISOR runs on a workstation and takes about 10 seconds per document to generate classifications and summaries. It is currently being applied to aircraft maintenance records and military intelligence messages.
>
> Fifteen programs similar to CONSTRUE and SCISOR were recently tested by the Defense Advanced Research Projects Agency for effectiveness in retrieving documents on a specific subject area (terrorist bombings) from 1400 documents (news stories). (Lehnert and Sundheim, 1991). Several techniques showed a good ability to extract and summarize relevant information from natural-language continuous text.
>
> Automatic message handlers, such as the Information Lens (Malone, 1987) and Lotus Notes, classify documents and automatically route them to those who need to respond to them. Citibank of New York, for instance, uses the IBS system (Sahin and Sawyer, 1989) to summarize, classify, and route telexes that they receive, even those containing grammar efforts and misspellings.

Many advanced text-management systems support hypertext (Nielsen, 1990). Hypertext is a way of providing portions of (possibly very large) sets of documents to readers. The document is generally composed of many relatively short "items" (a paragraph to a page), and the items are linked to facilitate cross-referencing. Readers can rapidly follow the links, which is much faster than thumbing through a printed document. For documents that are inherently linear, such as a novel, there may be only one relevant path between the items; following links would be like turning pages of a novel. But for other sets of metadata, such as information on how to use a computer system or information about the properties of an environmental data set, there is no single obvious path. Users follow paths to items appropriate to their particular questions. Context-sensitive help documents provided with many computer operating systems are examples of hypertext.

When hypertext technology is extended to include non-textual items, such as figures, audio, or video, it is called hypermedia. A textual description of an atmospheric acoustic sounder, with a link to a figure of an instrument and another link to the sound made by the instrument, would be an example of hypermedia.

2. Metalog: an electronic notebook for metadata

Metalog is an electronic notebook and a repository of information entered by the user and others. Metalog started as an electronic logbook for use in field experiments and data analysis. The system ran on workstations, but used a somewhat inflexible, command-based interface. Recently, the system was redesigned to be more flexible and easier to use. The new system, which has a graphical user interface, runs on Macintosh and PC-compatible (386 or more powerful) computers[1].

Metalog is like a database in that one can retrieve comments by any of several fields, such as type, project, and author. Metalog displays first lines of all comments retrieved, providing the user with a quick overview of what has been found. The user can then easily display and edit the full text of any of the found comments.

Unlike most databases, Metalog stores free-form text, and the text entries can be of any length up to 32,000 characters. And comments can be retrieved by any word or phrase in the text; retrieval is not limited by pre-defined keywords.

Metalog is like a word processor in that one can type comments directly into the system, can cut and paste them from other windows on the computer screen, and can load them directly from other text files. It differs from a word processor in that comments can be cross-referenced and listed in multiple directories. Thus, comments can be used to annotate other comments.

---

[1] Metalog and the COADS metadata are available to interested environmental scientists. The Macintosh and Windows 3 versions take about 1.8 megabytes of storage for the COADS metadata; empty Metalog (with help notes) takes about 0.4 megabytes. The ASCII-only text of COADS metadata is about I megabyte. Files are supplied on one or two diskettes. No runtime license fees are required. Requests should be directed to William Moninger, NOAA, Forecast Systems Laboratory, Mail Code R/E/FS1, 325 Broadway, Boulder, CO 80303; Phone: 303-497- 6435; Electronic addresses, Onmet: W.Moninger, Internet: Moninger@fsl.noaa.gov.

Because the system runs on personal computers, comments can be as private as the author desires. However, we have made provision for the sharing of comments between users. Comments can be printed or sent to a text file that can be read into other Metalogs or into any word processor document. Finally, comments can be sent to "metadata central." At present this is simply a file on each user's computer that appends all comments that are sent to it. The file can be periodically sent via mail or e-mail to a central location where the comments can be shared. Ultimately, we expect metadata central to be a direct link to one of several repositories of metadata.

In addition to its use in documenting COADS, which will be discussed below, Metalog is being used by several organizations. The Paleoclimatology Program of the National Geophysical Data Center is using Metalog to document a database on the little ice age, and to keep daily system records and instructions on how to use their graphical information system. The Wave Propagation Laboratory of NOAA and the National Center for Atmospheric Research have used earlier versions of Metalog to keep field notes for several experiments. Recently, copies of Metalog have been distributed to environmental researchers worldwide who have requested it. About half of the requestors are active users of COADS.

3. The use of Metalog in documenting COADS

During the past year, we have used Metalog to document COADS. The result is a version of Metalog that contains a great deal of COADS metadata. We call this COADS Metalog. COADS Metalog was released to the environmental science community in March 1992. We discuss here the process by which we produced COADS Metalog.

A small team of environmental and computer scientists performed the work. A "knowledge engineer" interviewed our senior expert on COADS and entered the knowledge gathered into Metalog. She and other team members also gathered existing COADS metadata for inclusion in the system. A programmer added capabilities to Metalog to support the COADS metadata. Other team members served primarily as system evaluators. In addition we received guidance from an advisory committee of environmental and computer scientists.

3.1 Design decisions

We decided to target as primary users of the system experienced environmental science researchers who are not familiar with marine in-situ measurements. That is, we assumed that users would be familiar with the rudiments of statistics, and have a level of mathematical sophistication appropriate to advanced graduate students in the natural sciences, but would have no familiarity with spatially and temporally varying data.

We also desired that the system be useful-—or at least credible—to current users of COADS. If this group didn't support COADS Metalog, it is unlikely that inexperienced users would trust the system enough to use it.

We decided to support all of the unstructured textual metadata discussed in section 1.2. This would include relevant refereed journal articles and abstracts, and gray literature such as internal reports and informal comments.

In some conflict with this, we also desired that COADS Metadata have as much credibility as possible. Thus, everything that appears in COADS Metadata was checked for accuracy by our COADS expert. Moreover, when items of metadata already existed in printed form, we tried to ensure that the electronic version included in COADS Metalog matched the printed version word-for-word. (It was not possible to achieve an exact match in appearance because Metalog does not support multiple fonts.)

We finessed potentially troublesome issues of intellectual property and copyright in two ways. First, through a fortunate choice of authors: the journal article and technical report we included in COADS Metalog are works of the U.S. government, and hence are not covered by copyright. Second, by the doctrine of "fair use" (Jones, 1987): portions of documents that would otherwise be covered by copyright (e.g., abstracts) could be included because their use is restricted to research, and COADS Metalog is designed as a research tool.

We chose to use hypertext, because it is efficient at providing small, relevant portions of large sets of documentation to users.

We chose to develop the system initially to run on Macintosh computers and PC's, because we believe these are more readily available to environmental researchers world wide than are workstations.

We chose not to include structured metadata initially. This kind of metadata is readily supported by currently available data base management systems; we felt it would be more valuable to address the less-well-studied question of unstructured metadata.

Finally, we chose to postpone support for hypemedia, in particular, figures, equations, and multiple fonts, in favor of focusing on ease of entry and support for cross-referencing of textual information. Non-textual items do not yet have well-accepted standards for interchange among different computer platforms, and supporting these items on any one platform takes a considerable programming effort.

3.2 Contents

COADS Metalog as released in March 1992 contains 1.8 megabytes and consists of 628 comments and directories. Figure 1 shows the top-level structure of the contents; figure 2 shows an expanded view of one of the directories. Individual comments can and do appear in as many directories as are relevant. Contents include the following:

- A set of comments that form a tutorial on how to use Metalog to access the COADS metadata.
- Thirty-nine comments on specific aspects of COADS derived from interviews with our COADS expert.
- Fifty-five comments that consist of the text of a journal article and a technical report about COADS (Woodruff et al., 1987; Slutz et al., 1985).
- More than 50 abstracts of journal articles relevant to COADS found through a literature search and downloaded from an abstracting service.

- More than 200 citations relevant to COADS from a climate researcher's personal database of papers and talks.
- Sample information on routes and instrumentation for 10 ships, generated from one of the World Meteorological Organization data tapes (WMO, 1955-).
- Abstracts of 38 talks given at the 1992 International COADS workshop.
- One comment that is the text of an article about the digitization of historical data for COADS. (Elms, et al., 1992)
- Ninety-five cross-referenced directories of comments.
- Eighty-four comments providing help on how to use Metalog. (These are in a separate Metalog from COADS Metalog, but are included in the distributed software.)

In addition, COADS Metalog has space for users' comments. These can annotate the information already in the system or express other ideas. Also, users can develop their own directories and hierarchical structure. The number of additional comments and directories is limited only by imagination, enthusiasm, and disk space.

4. Lessons learned

Metalog is helpful in documenting environmental data. Users report that field notes taken with Metalog are more complete, secure, and easier to distribute than notes handwritten in logbooks. Users at the National Geophysical Data Center and our own experience in documenting COADS indicate that Metalog makes it easier to document existing data.

Metalog allows efficient maintenance of documentation. In the final months before release of COADS Metalog, only a limited amount of staff time was available to maintain and expand the COADS metadata. A non- technical staff member not particularly comfortable with computers was able to use Metalog effectively to add and cross-reference items of metadata, with minimal supervision. A scientist was able to download items of metadata from electronic networks quickly and easily.

COADS Metalog provides information that COADS researchers believe will be useful. When COADS Metalog was demonstrated at the 1992 International COADS workshop, approximately 20 researchers requested copies. (These researchers have just received their copies, and we have not yet received feedback from them.)

Our design decisions appear to be appropriate, based on feedback from our advisory committee. In particular, our emphasis on content over form (i.e., support for ASCII text only) in this early stage of development was endorsed. In addition, our decision to build in a high level software environment (Goodman, 1988) rather than to attempt to buy ready-made software was approved.

Figures and "rich text" (equations and multiple fonts), while clearly necessary in the long term, do not appear to be essential for initial use of the system. Their absence has caused few if any problems in field use of Metalog.

Authority of comments is important. Readers feel a strong need to know the source of items of metadata—whether it is from a journal article or from an off-the-cuff remark. Comments that are written, edited, and corrected by several people need to have a well-defined author history so that thoughts are not incorrectly attributed. Comments that are incorrect, or correct but naive, can have a strong negative impact on the credibility of the entire package of metadata.

The integrity of authors must be respected, or they may become hostile to the entire idea of electronic data documentation. Comments taken from journal articles should be verbatim from the printed version. in fact, we would not include a journal article in Metalog without approval of the author, simply because the difference between ASCII text and the typeset text might cause offense.

Many comments about data do not need a reference to a particular time and place. Initially, we believed that all comments would need such a reference, inasmuch as they are comments about data, and data is only meaningful for a particular time and place. We quickly found that, at least for COADS, most comments refer to the data set as a whole, or to large spatial or temporal segments of the data. It makes no sense to give a precise spatial and temporal reference to a comment such as “Long-term trends in wind-speed may be artifacts because A structure (such as we originally anticipated) that required such references for every comment would be unnecessarily cumbersome.

5. Future directions

5.1 Metadata

We have requested that all users of COADS Metalog use the system to freely annotate the existing metadata and add their own comments. We have asked that they send comments that seem to be relevant to the COADS research community to us for inclusion in the next release of COADS Metalog. We will follow up this initial request with another request and possibly with a piece of software that makes it easy for users to identify their new comments and write them to a file for delivery to us.

On the basis of this information, and on new information that we gather from other sources, we will correct and expand the COADS Metalog and release a new version.

Now that we have developed and begun to hone our techniques for assembling electronic data documentation via COADS, we intend to use Metalog to document additional environmental data sets. In addition, the Metalog development team is available to others who wish to use Metalog to document their own data sets.

5.2 Metalog

Further development of the Metalog software will be dictated in large part by feedback from our users. We have recently provided Metalog to approximately 50 users, and we will actively solicit their views on how the system can be improved. On the basis of feedback we have already received from potential users, we anticipate doing the following within the next year or two:

- Develop the capability of semi-automated comment classification, cross-referencing, and retrieval using artificial intelligence and other software technology.
- Develop version management techniques that integrate scientists' private comments with new, centrally distributed metadata, while maintaining appropriate cross-referencing.
- Develop software that makes it easier for scientists to share metadata via electronic networks.
- Develop a faster PC version of Metalog.
- Port Metalog to run under the X-windows user interface environment.
- Increase the richness of metadata supported by adding figures, equations, multiple fonts, and possibly audio.
- Integrate Metalog with other systems, such as the NOAA Earth System Directory and structured databases.

6. Conclusion

We have demonstrated that computer support for textual information can be helpful in documenting a large and widely used environmental data set. The system is effective in that it provides needed information readily, and it is cost-effective in that it can be used by data clerks as well as scientists. We believe Metalog, and the general category of systems similar to Metalog, offer benefits to individual users and to scientific communities.

For individual users as readers, Metalog-like systems offer rapid access to a broad spectrum of metadata, from journal articles to informal comments. And they provide effective retrieval of relevant metadata.

For individual users as writers, Metalog-like systems provide an organized repository of comments that can be private, shared with a small work group, or shared widely. The comments can be copied and archived easily, so they are at less risk of loss than comments written in personal notebooks.

For scientific communities, either informal communities of researchers sharing similar interests or formal communities of coworkers such as NOAA's laboratories, Metalog-like systems offer the following:

- Increased productivity as scientists use electronic metadata to make better use of existing data sets.
- Decreased cost as secretaries, data clerks, and scientists create and manage metadata more efficiently.
- Increased esprit de corps as scientists share information more readily.
- Increased credibility as knowledge assets are managed more effectively. The metadata stored and made available in Metalog-like systems can be used as the basis for policy papers to inform the government, educators, and the general public about environmental phenomena.

7. Acknowledgments

For early advice and guidance, and provision of extensive information from his personal database, we thank Klaus Wolter. For actively using Metalog in one or more of its several incarnations, and providing helpful feedback, we thank Leah Eicher, Cecilia Girz, Wendy Gross, Charlie Martin, and Jon Overpeck. Ray McCall provided useful advice from his perspective as an experienced hypertext researcher, and much appreciated moral support. This work was funded in part by NOAA's Climate and Global Change Program office.

8. References

Elms, J.D., Quayle, R.G., Woodruff, S.D., 1992: Digitizing historical merchant marine weather observations. Submitted to *Mariners Weather Log*.

Hayes, P.J. and Weinstein, S.P., 1990: CONSTRUE/TIS: A System for Content-Based Indexing of a Database of News Stories. *Proceedings, Second Conference on Innovative Applications of Artificial Intelligence*. American Association of Artificial Intelligence, Menlo Park, CA. pp. 1-5.

Goodman, D., 1988: *The Complete HyperCard Handbook*, 2nd Ed., Bantam Books, New York, 875 pp.

Jacobs, P.S. and Rau, L.F., 1990: SCISOR: A system for extracting information from financial news. *Communications, Association for Computing Machinery*, **33**, 88-97 (November).

Jones, H.W., III, 1987: Developing and Distributing Hypertext Tools: Legal Inputs and Parameters. In: *ACM Hypertext Compendium, Release 1.0 for Apple Macintosh* (1991). (ACM order number 217913.)

Lehnert, W., and Sundheim, B., 1990: A Performance Evaluation of Text-Analysis Technologies. *AI Magazine* **12**, 81-94 (Fall).

Malone, T.W., Grant, K.R., Turbak, F.A., Brobst, S.A., Cohen, M.D., 1987: Intelligent Information Sharing Systems. *Communications, Association for Computing Machinery*, **30**, 390-402.

Moninger, W.R., Ignatowski. A., Fozzard, R., 1990: Progress report on Metalog: A System to capture and maintain environmental metadata. Preprint volume, *Sixth International Conference on Interactive Information and Processing Systems for Meteorology, Oceanography, and Hydrology.* AMS, Boston. pp 179-186.

Nielsen, J., 1990: *Hypertext and Hypermedia*. Academic Press. 263 pp.

Sahin, K., and Sawyer, K., 1989: The Intelligent Banking system: Natural Language Processing for Financial Communications. In: *Innovative Applications of Artificial Intelligence* H. Schorr and A. Rappaport, (eds.), AAAI Press, Menlo Park, CA. 363pp.

Slutz, R.J., Lubker, S.J., Hiscox, J.D., Woodruff, S.D., Jenne, R.L., Joseph, D.H., Steurer, P.M., Elms, J.D., 1985: *COADS, Comprehensive Ocean-Atmosphere Data Set, Release 1*. ERL Special Report. Available from NOAA Climate Research Division, ERL, Boulder, Colorado.

WMO (World Meteorological Organization), 1955: *International List of Selected, Supplementary and Auxiliary Ships*. WMO No. 47, Geneva, Switzerland.

Woodruff, S.D., Slutz, R.J., Jenne, R.L., Steurer, P.M., 1987: A Comprehensive Ocean-Atmosphere Data Set, *Bulletin of the American Meteorological Society*, **68**, 1239-1250.

Woodruff, S.D., Lubker, S.J., Quayle, R.G., Radok, U., and Doggett, E.D., 1991: *Differences Within and Among Surface Marine Datasets*. NOAA Environmental Research Laboratories, Climate Research Division, Boulder, Colorado, 216 pp.

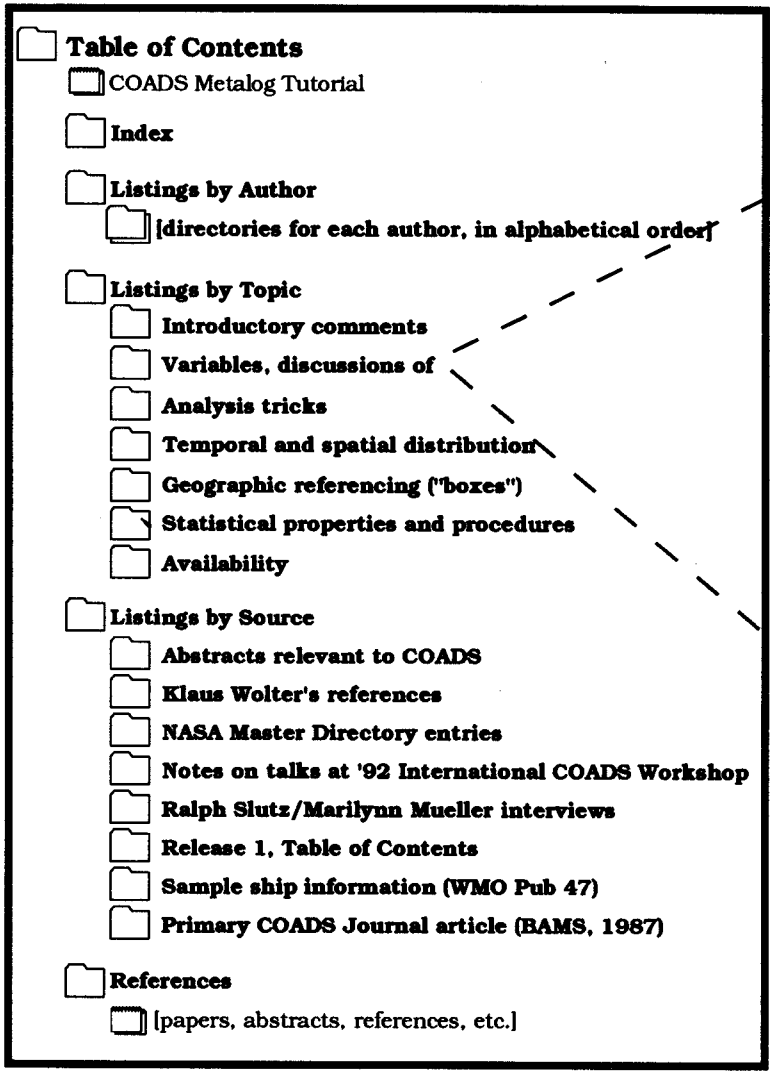

Figure 1

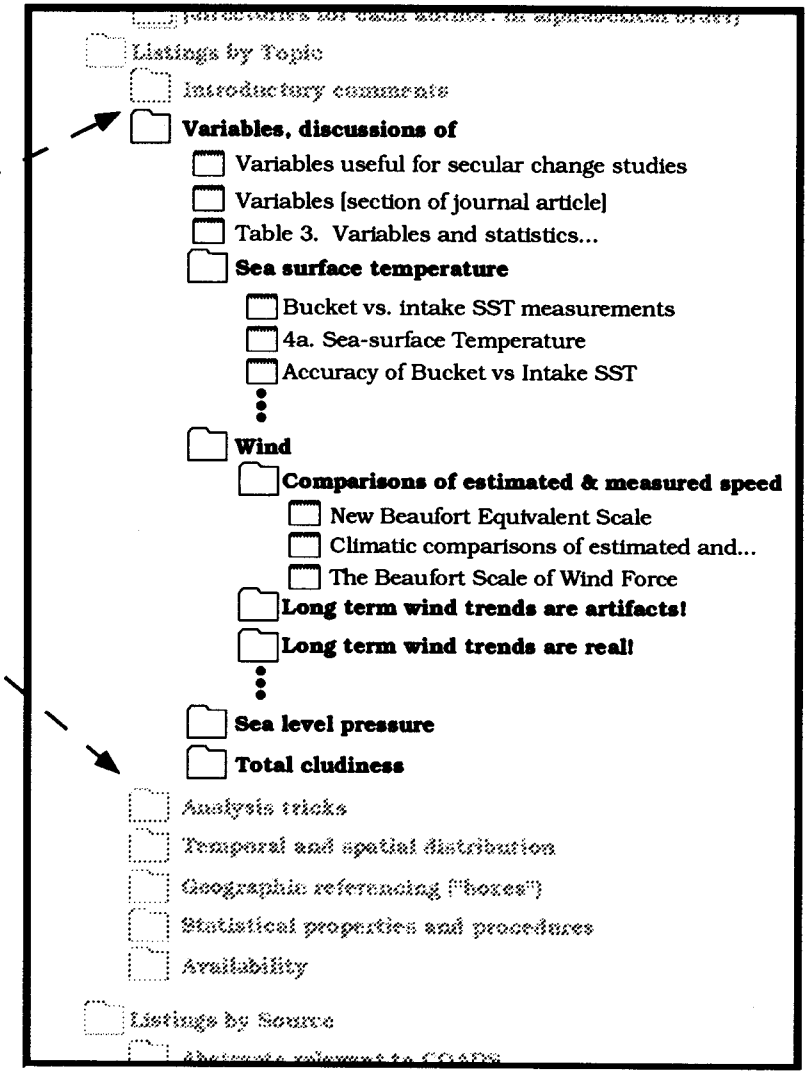

Figure 2

Legend: **Directory** Comment