

ISI's Implementation of Network Resources in Secondary Databases: Selection Criteria, Metadata, and Linking

Yasuki Kaneko

ISI -Thomson Scientific-

Palaceside Bld. 5F, 1-1-1 Hitotsubashi, Chiyoda-ku, Tokyo 100-0003, Japan

Abstract

The importance of information resources on the Internet is growing rapidly. In the academic field, the latest information is communicated over the Internet.

In this article, it is discussed how one should select scholarly web sites as content for secondary databases, how the bibliographic information in databases should be described, and how links between electronic information should be created.

Keywords: *Web Site, Selection Criteria, DTD, Algorithmic Link*

1. Introduction

Over the past few years, the volume of scholarly information available over the Internet has increased dramatically and rapidly. The Internet has become an important resource to officially announce research results faster, as well as to retrieve the latest information.

There are many advantages to using the Internet for research:

- It is possible for researchers to communicate easily, enabling more joint research projects regardless of the researchers' locations
- It is possible for researchers to obtain the latest information easily

There are also disadvantages such as:

- It is more difficult to find the most relevant information because the volume of information has increased greatly
- There is an absence of the peer-review process with web documents

A bibliographic database is a useful tool to search

for information efficiently. A bibliographic database exists to provide guidance in finding necessary information; in that regard, it is required that the data is organized optimally.

One way to maximize the efficiency of a database is for the database editor to choose a minimum of bibliographic elements in order to keep response time short.

Now that the Internet is becoming a mainstream media source, it is now expected that secondary databases will include Internet resources in their source lists. When adding information from the Internet, in addition to the traditional books and journal articles, it is important to determine the necessary bibliographic elements and their syntax.

2. Selection Criteria

The Internet and WWW technology allows people all over the world to post and retrieve data easily. As a result of this, all kinds of information (e.g. personal interests, advertisements of companies, research results, commodity exchanges, governmental official opinion, large-scale statistics, etc.) now exists on the Internet.

Reliability of academic information has been regarded as questionable since the early days of the Web when researchers and librarians began to pay attention to the Internet as an information resource. Unlike the peer-review system with academic journals, there are no similar filters for the Internet. Therefore it becomes necessary to judge whether that information is scientifically trustworthy.

A general web search engine can retrieve

information from all over the world without regard to a robot-type or a category type. However hit counts will be vast, and it is difficult to find valuable information from the results.

In academic libraries it is the librarian, in many cases, who performs these filters. The librarian actually checks the Web site. The reliability of the information is then examined and only valuable information is added to the links collection.

For over 40 years, ISI has consistently and carefully selected the information it provides to users. This same consideration for selection is applied to Web documents which recently have been added to the coverage of Current Contents Connect®.

Although the selection criteria for Web documents generally follows the journal selection criteria, it is also necessary to create original criteria for selecting Web documents.

ISI adopts many criteria such as Authority, Accuracy, Currency, Navigation & Design, Applicability & Content, Scope, Audience Level, and Quality of Writing as selection criteria of the web sites and documents.[1]

3. Metadata for the Web Site

3.1. Types of Metadata

Metadata is “the data about the data”, or it can be defined as “the set of the attributes and elements which are substituted for the target data”. There are various types of metadata, depending on the object type. For example, in addition to the Dublin Core, the following metadata categories exist:

- **Metadata for Social Sciences Datasets**

Metadata for social sciences data is usually called a “code book”. It is not only a description of the outline of the data, but also an explanation of how each investigation is done, the length of each field, and the meaning of the variables contained therein. The most famous metadata for social sciences data is the DDI from ICPSR[2].

- **Metadata for Geographic Information Systems**

The Geographic Survey Institute of Japan explains the metadata for GSI:

It points at the data which describes the whereabouts of the space data (geography information), contents, quality, use conditions, and so on. "The information which is necessary to use information" can be shown separately from the geography information. The user of the geographical information examines these metadata, and judges whether that geographical information is available. Therefore, it must be described to the metadata with sufficient common information such as "Where can you use it if it exists by what kind of shape and does what?". [3]

3.2. Metadata for Current Web Contents

ISI creates metadata for individual web sites in Current Web Contents(CWC). Listed below are the descriptive elements[4].

According to the classification of Dublin Core, which was given by Hillman,[5] bibliographic elements in the CWC are grouped as below.

- **Contents**

Web Site: The title of the Web site as defined by an ISI subject editor. This could be the page title as designated within the HTML coding or the full text title appearing on the page. The URL for the Web site is also shown here in its entirety.

Description: A concise description of the entire Web site highlighting major content features. This element is written by an ISI subject editor.

Keywords: A list of terms that describe the subject specific nature of the Web site's contents. Keywords are created by ISI subject editors.

Type: The nature or specific kind of information contained on the site.

- **Intellectual Property**

Author: The creator of the site as identified by an ISI subject editor according to information contained within the site.

Publisher: The organization that is identified as the copyright holder for the content of the site.

Provider: General description of the sponsoring institution(s) for the Web site.

- **Instantiation**

Language: The language(s) used in the site.

Format: Specific electronic formats used on the Web site. Links checking is a separate activity that is performed on a daily basis.

ISI also shows the date on which the site was last reviewed by an ISI subject editor.

3.3. Web documents

ISI selects web sites for coverage in Current Contents Connect. After selecting the web sites, ISI subject editors examine all “child” URLs of the selected web sites. Approximately 40% of the reviewed “child” URLs contain full text documents that are included in the ISI Current Contents eSearch collection. Metadata descriptions are assigned to the web documents to enable efficient search capabilities; this precludes the need to search the full text of each web document and can save the researcher valuable time.

4. ISI LINKS

The biggest advantage of data on the Web is its direct access to related information via the link function. ISI also creates links from secondary databases to primary text.

ISI currently receives the necessary information from the publisher and creates a link. This information is called “feed”. Figure 1. is a sample DTD document.

Figure 1. Sample DTD (Document Type Definition)

```
<ArticleSet>
  <Article>
    <PublisherId>HW_sci</PublisherId>
    <ArticleId>sci;287/5454/811</ArticleId>
```

```
<Journal>
  <PublisherName>American Association
for the Advancement of
Science</PublisherName>
  <JournalTitle>Science</JournalTitle>
  <Volume>287</Volume>
  <Issue>5454</Issue>
  <PubDate>
    <Year>2000</Year>
    <Month>Feb</Month>
    <Day>04</Day>
  </PubDate>
</Journal>
<ArticleTitle>PALEOBIOLOGY:Reef
Processes in the Long View</ArticleTitle>
<FirstPage>811</FirstPage>
<LastPage>812</LastPage>
<AuthorList>
  <Author>
    <FirstName>J. E. N.</FirstName>
    <LastName>Veron</LastName>
  </Author>
</AuthorList>
<FullTextURL>http://www.sciencemag.org/
cgi/content/full/287/5454/811</FullTextURL>
</ArticleSet>
```

It is also adaptable for an Open URL syntax.

ISI is now developing an algorithmic link system. It will soon be possible to create and maintain links algorithmically. Therefore it will be possible to make links easier and faster.

5. Conclusion

The Internet has become an important resource for efficient information retrieval and as a data source. There is a lot of relevant content on the Internet that, when indexed, can be searched efficiently. The importance of selecting such web documents was discussed, and the necessity of using a DTD for links was also discussed. Selection criteria, metadata and links all work together to enable the creation of a seamless, comprehensive environment for researchers.

Acknowledgements

I would like to offer my appreciation to Jeff Clovis, Philip Heller, and other ISI staff for providing information.

References

- [1] Testa, James. Current Web Contents : developing web site selection criteria. [Online] [<http://sunweb.isinet.com/isi/hot/essays/selectiofmaterialforcoverage/23.html>]
- [2] Data Document Initiative. [Online] [<http://www.icpsr.umich.edu/DDI/>]
- [3] Geographic Survey institute. What is a metadata?. [Online][Japanese]. [<http://zgate.gsi.go.jp/keisiki.htm>]
- [4] [Institute for Scientific]. Current Web Contents Record. [Online] [<http://sunweb.isinet.com/isi/products/cc/cwc/biblio.html>]
- [5] Hillmann, Diane. Using Dublin Core. Apr. 12, 2001. [Online] Available: <http://www.dublincore.org/documents/2001/04/12/usageguide/#coreelements> [Accessed Jun. 27, 2001]