A

ITSCG AWG 1 N27
ITSCG AWG 2 N20
September 1995
Originator: ISO/CS IPS

The Discovery and Retrieval of Internet Resources from ISO

1. Description of the problem

ISO provides information and documents from many sources: the ISO/CS WWW server, FTP and WWW servers of technical committees, etc.. It is difficult for users to find, browse, and retrieve information without knowing the address of these servers. In addition every server has a different interface and structure which complicates efficient browsing. Finally, there is no method of sending a query to search through the contents of all the servers in question.

A request has been made by JTC 1 to the ITTF for the Central Secretariat to provide a coordinated service by having hyperlinks from an ISO/CS WWW page to the JTC 1 subcommittee servers on the Internet. This paper proposes the *ISO Web*. *ISO Web* will comprise of all ISO standards-related servers distributed throughout the Internet. *ISO Web* will provide a consistent interface and a content indexing system that conserves network bandwidth (i.e. reduces unnecessary network traffic).

2. Requirements for the *ISO Web*

Some requirements for this new Web are as follows:

1. The service shall be distributed. Documents are maintained by the subcommittees and shall be made available on their own server.

2. The service shall provide browsing capability. A user shall not need to know multiple number of addresses. Given the address of any one server on the *ISO Web*, the user shall be able to browse throughout the whole *ISO Web* using hyperlinks only.

3. The service shall provide the capability to search for documents or information. Indexing of server content should be performed by the committee or working group in charge of the server. If expertise and

resources are unavailable for a particular group, then the indexing of that server's contents shall be performed by the Central Secretariat.

4. The service shall only provide access to authorized users.

3. Possible implementation

The following is a possible implementation.

3.1. Registration procedure for servers

A registration procedure is needed in order to confirm that the server is the official server that represents a particular committee. The registration information should include:

the name of the subcommittee
the name of the contact person or server webmaster
the e-mail address of the contact person
the URL (Uniform Resource Locator) of the server

An alias for the Internet domain name of the host used for the server is recommended. Normally, the alias has "www" followed by the sub-net name, for example, www.name.xx where "xx" is the top-level domain. The host machine that the server runs on can be changed, while the alias remains the same.

This information should be sent to the representative at the ISO/CS (ITTF). The registration can be done via a paper form or by a Web Form. An example registration Web Form can be found at

http://www.iso.ch:8080/isob/ISOregister

3.2. Proposed structure of the *ISO Web*

A minimal structure of the Web is necessary to satisfy Requirement 2. The structure should be discussed with the various subcommittees in order to satisfy the needs of the community. A possible implementation that has the minimum number of constraints is given below. Note that the use of JTC1 in the URLs below are examples only.

1. *ISO Web* home page
This page will provide the hyperlinks to all registered servers of technical committees, subcommittees, and working groups.

2. TC server home page
The URL of this page should be of the form

http://www.name.xx/JTC1

This page should contain hyperlinks to the home pages of its SCs and WGs and to the *ISO Web* home page. Any documents available on this server should follow guideline 5 and 6 below.

3. SC server home page
The URL of this page should be of the form

http://www.name.xx/JTC1SC1

This page should contain hyperlinks to the home pages of its WGs, to the TC server home page if one exists, and to the *ISO Web* home page. Any documents available on this server should follow guideline 5 and 6 below.

4. WG server home page
The URL of this page should be of the form

http://www.name.xx/JTC1SC1WG1

This page should contain hyperlinks to the SC server home page if one exists, to the TC server home page if one exists, and to the *ISO Web* home page Any documents available on this server should follow guideline 5 and 6 below.

5. Structure or "path" to the documents
The following structure is based on the ISO policy document, "Policy Concerning the Distribution of ISO Documents Electronically". The home page of a server should have 3 hyperlinks to the following pages:

http://www.name.xx/JTC1SC1/publications
http://www.name.xx/JTC1SC1/projects
http://www.name.xx/JTC1SC1/administration

This structure is recommended in order to simplify the enforcement of different security requirements for each type of document. In addition there is the added advantage that separate indices can be created automatically for documents of a particular type.

6. Documents uniquely identified
From each of the three document type pages, the subcommittee may provide hyperlinks to their documents. The option of adding additional levels in the "path" is left to the choice of the maintainer of the server. However, the "path" should be chosen carefully and once it is created, it should be considered frozen. The reason for this is that the document may be referenced from another document. Documents should be replaced, but not removed. The problem of URLs which confuses the name and the location of a document will be addressed by the URI (Uniform Resource Identifier) Internet working group. Until the Internet standard is released and implemented, the document should be uniquely identified based on guidelines for the identification of documents as provided by ITSCG AWG 1.

Note that it is possible to have more than one home page on one server. For example, the JTC 1 home page and the JTC 1 SC 1 home page may exist on the same server, while the JTC 1 SC 2 home page may be on another. It is also possible that intermediate home pages may not exist. For example, TC 1 SC 1 may run their own server, but a TC 1 home page may not exist as of yet.

3.3. Contents of the *ISO Web*

All documents of the *ISO Web* will be entered by the committees and working groups. The groups are free to enter what is needed, subject to the ISO policy restrictions. The control of presentation, such as the use of graphics and HTML tags, is also left to the groups, subject to some technical guidelines:

1. The current standard of HTML, as released by the HTML Working Group, shall be used. Extensions shall be avoided. Extensions tend to fragment the Web and their use is not good practice as they are often browser-specific.

2. The content of elements such as <TITLE>, <CITE>, and <ADDRESS> are given certain attributes in the indexing mechanism as described in section 3.4. In this way it will be possible to query only within titles, within references, etc. Therefore consistent use of these element types will allow users to find your documents more readily and create a more powerful search service.

All HTML documents shall have a <TITLE> element. All references shall be contained within <CITE>. Information concerning the author of the document shall be contained within <ADDRESS>.

3. It is expected that many documents will be made available in formats other than HTML, SGML, or plain text. Normally this would pose a problem in indexing for documents where the text is unavailable such as TIFF or other image files, audio files, or proprietary binaries such as Word. If documents are only available in such a format, then a complementary abstract document in HTML shall be made available with a hyperlink to the non-text document. This would allow a summary of the contents to be indexed. There are some notable exceptions to this guideline and they are described in section 3.4.

3.4. Indexing of the contents of the *ISO Web*

In order to satisfy Requirement 3, an index of the content of the *ISO Web* is needed. Due to its distributed nature, some additional issues must be addressed that are not normally encountered while indexing a single server. To preserve network bandwidth, indexing should be done locally and the use of software "robots" should be avoided. However, a TC, SC, or WG should be allowed to choose whether they prefer to index their own contents or to let another server index their content for them.

One public domain tool exists that provides a solution. The package, called Harvest, is the result of an IRTF (Internet Research Task Force) project. The advantage of a public domain tool is that this will satisfy the JTC 1 requirement to not rely on any particular vendor. Harvest is a powerful set of integrated tools that can gather, extract, organize, search, cache, and replicate relevant information across the Internet. The system allows for the creation of local indices that are exported automatically to another server, but also allows for indexing over the network. The advantages of distributed indices cannot be understated. With the Harvest system, one can select the indices that would form the basis of one's querying system. For example, the JTC 1 server may provide a WWW form to query only the contents of its own SC's and WG's servers, while the *ISO Web* home server may provide a form to query all servers in the *ISO Web*. To preserve network bandwidth even further, the Harvest system provides a replication mechanism (often called "mirroring") of its indices. This allows one querying system to be copied to another server. Thus the gathering of indices has to be done only once with the ability of querying being available on different servers.

Normally documents in formats such as postscript, RTF, or Framemaker MIF will require an abstract document in order to summarize its content for indexing. The Harvest system, however, provides a special set of "summarizers" that extract data and generate a content summary from these formats so that separate abstract documents are not needed.

Finally, the Harvest system is not limited to the indexing of WWW resources. It can gather and index files via other protocols such as FTP, Gopher, and NNTP (Usenet news).

In summary, Harvest provides a flexible indexing and querying solution that is decentralized, conserves network bandwidth, and is not specific to any vendor, application protocol, or format.

3.5. Security of services

The use of authentication mechanisms is needed to satisfy Requirement 4. Note that as the documents are under the control of the TCs, SCs, and WGs, it is the server maintainer's responsibility that proper levels of security are provided on their server for those documents that require protection. Most http servers provide the ability to apply the Basic Authentication Scheme. This scheme is secure enough for documents of minor importance or commercial value. However, it may be insufficient for official publications or project-related documents. Some recommendations for these documents will be provided by ITSCG AWG 2.

4. Pilot project proposal

In order to test this implementation, a pilot project is proposed. The ISO Central Secretariat with the help of two or three subcommittees can prepare the *ISO Web* to test its feasibility. JTC 1 SC 22, JTC 1 SC 24, and TC 184 SC 4 are possible candidates for participation in the pilot project as they already maintain servers. Their servers are located in Australia, the Netherlands, and the US respectively. The widely distributed servers can test the capacity of the network and the performance of the system.

A prototype *ISO Web* can be found at the following address:

http://www.iso.ch:8080/tmp/ISOWeb.html

Note that there are three subcommittee servers and 12 documents are indexed in this prototype.