

# **The Strategy to Organize Electronic Documents on Internet from the library approach**

**Ya-Ning Chen**  
**Computing Centre, Academia Sinica**

## **Abstract**

By application of Internet and electronic publication, it has become one of most important and hottest topics to find out an effective way to organize electronic documents on networks. Due to lacking of irreasonable organization of these documents, it is very difficult for users to (Yet the appropriate information at right time, especially these documents were put on servers and accessed through network's connections. As a result, many diverse experts at different disciplines have been practiced research programs to fix this problem through other existed Internet navigators, for example Archie, WAIS (Wide Area Information Server), Gopher and WWW (World-Wide Web), etc. It also be-ins to be noticed this problem and tries to offer the solutions in the Library field, because it is known for organizing the information effectively to help user getting the right one. This article is to discuss how to organize electronic documents by following key points: changes to the bibliographic control environment. analysis of the Chinese electronic documents, standards and derived problems in practice, noticeable trends, suggestion and conclusion.

## **1.0 Introduction**

By advancement of the Internet and its extended application, it becomes another precious treasure to die, out information. Though many existent Internet navigators are to help users finding information, but the results aren't always matched to their needs. This article tries to discuss how to organize electronic documents on networks from the library's bibliographic control viewpoint and offer a practicable solution.

## **2.0 Changes to the bibliographic control environment**

What is the difference to the bibliographic control between electronic documents and printings ? Before organizing these e-documents, first we have to clarify what changes occur to the cataloging environment. They are set out by these points as below.

### **2.1 Catalog's function: from finding and gathering to identifying, access and directory gateway**

According to Cutter's presentation, the catalog has two main purposes to enable a user to ascertain if the library has a particular item and the library catalog should show what items the library has that share a common characteristic, that is finding and gathering<sup>4</sup>. Now Internet has expanded the catalog to identifying, access and directory gateway. Not only can netusers retrieve the bibliographic information, also the full text can be accessed via network's connections. Moreover, documents have different derivatives both in electronic and printing format. Finally some networked resources can be accessed only by the specific

---

4 Robert H. Burger, *Authority Work (Littleton Colorado : Libraries Unlimited, Inc., 1985), p.4.*

person or group, and catalog has to indicate the relevant information telling this kind of restrictive use.

## **2.2 Cataloging scope: from local and physical to global and virtual**

Traditionally, the library only organizes materials that it owned. Via Internet, there's no distinction to access physical and virtual information. Indeed it means that user can retrieve any local and remote information simultaneously no matter where it is located. Therefore the cataloging scope is transformed from local to global and from physical to virtual.

## **2.3 Data : from single medium to multimedia and from linear so hyperlink**

Currently most of information carriers are printed formats. But this trend has been altered gradually with the application of digital technologies and its derived products. We can find one fact that the data is composed of more than two kinds of media, that is, the medium of information carriers will become multimedia, especially to the digital documents on Internet. For example, the WWW's page is a mixture of image, graphic, sound and motion picture. In addition to medium mposition, the arrangement also has different changes. Traditionally, data is arranged in a linear way with page number. With the appearance of WWW, this kind of arrangement has also been transformed to be hyperlink, a kind of Jumping pointer to the related documents directly without going through document page by page. Therefore not only have the above changes been affected the type and representation of information carriers radically, but also the organization has been affected.

## **2.4 Retrieval result: from bibliographic and holding record to full text**

The information that user gets is full text, not just bibliographic and holding record through OPAC's retrieval as before. On Internet the users can get the bibliographic information first, then access to them without visiting to library to lend it out. For example, we practiced a search by WWW's search engines (ex. InfoSeek, Lycos, etc.) and these engines give us a list of bibliographic information with information title, summary and URL (uniform resources locator). By following to the URL linkage, then we get the original document. In the near future, the tendency to get full text will also affect the catalog and its process.

## **3.0 Analysis of the electronic Chinese documents**

According to my personal survey on the electronic Journals and the Chinese electronic documents of two TANet's FTP sites (moers2.edu.tw, nctuccca.edu.tw), there're some interesting findings that I found. Though these documents are only small partial of Internet resources, the findings also reflect some characteristics of networked resources. These findings are described briefly as below<sup>5</sup>:

### **3.1 More than two file formats**

The file formats existed on Internet are more than two types. Besides the Plain ASCII, there're existing Tex, Postscript, etc. And their resolutions are also different. Sometimes the same file format has diverse versions with different resolutions. For example, the CERFnet News, NEARnet Newsletter, Solstice, Tidbits are all of Postscript file, and then the Ghostscript

---

5 Ya-ning Chen, *Discussion on Organization of Network Resources from Cataloging Viewpoint Taipei, Taiwan : Wen-Hwa Management Information Inc., 1995*, pp. 133-151

program can't read them out correctly due to the various format versions. On TANet there're only three kinds of file formats, that is the Plain ASCII, Postscript and GIF.

### **3.2 Inconsistency in position of bibliographic information**

The bibliographic information appears in the middle of e-documents from the Project Gutenberg offerings. But as for the e-journals and e-newsletters<sup>6</sup> are at bottom. All of samples on TANet are consistent in appearing at top. As for the position of bibliographic information, the e-documents have no same rule to this point.

### **3.3 Difference in e-document's structure**

No consistent rules is made to e-document's structure both in English and Chinese. The three parts, title page, introduction, and table of content (TOC), are components of edocuments in Project Gutenberg. The e-Journals and e-newsletters are similar to the printings. In addition to presenting the above parts, they also offer the accessing method, for example the Public-Access Computer Systems Review. But there's one special point to the Bryn Mawr Classical Review, it only offers the content without anything such as title page, etc. On TANet, the Chinese e-documents have similar distinctive as the English e-documents.

Classical Review, it only offers the content without anything such as title page, etc. On TANet, the Chinese e-documents have the similar distinctive as the English e-documents.

### **3.4 Not all of networked resources with introductory**

In cyberspace the networked resources are accessed by FTP, Gopher or WWW. Not all of them offer the introductory information to indicate what they are. For example, the FTP sites always set up the README file as an introductory to the site, its directories and files. As well as the Gopher and WWW.

### **3.5 Diverse functions of README file on TANet**

From my survey, it revealed the README files of FTP sites on TANet have three functions TOC, access and usage. The TOC only explains the resource's title, file name and content, but access also offers more information telling the netuser accessing way to get the file and author's brief introduction. Finally, the usage function gives more detailed information how to install and use the program software. Furthermore, there's an interesting phenomenon about the function among the README, README.chinese and INDEX. The README is an explanatory guide how to FTP the file in English, as for the README.chinese in Chinese. Moreover, the INDEX is a map of indicating how many files are saved at FTP sites. Therefore, we can conclude that the three files have the same or similar function. The Gopher and WWW also have the similar set up. For example, the Gopher has "About ..." or "Browse..." option in Menu, and as homepage is to the WWW servers.

---

6 Michael Strangelove, *Directory, of Electronic Journals and Newsletters* (Ontario, Canada : University of Ottawa, 1992).

FTP : ' ftp csuvox1.csu.murdoch.edu.au  
at/pub/library/e-journals.dir directory

### **3.6 Unconformity between the README and the related files**

The README file of FTP site is to indicate what the site contained and saved, so it is a indicative pointer. Judged from the NCCTU (nctucca.edu.tw) site, I found the README file isn't updated currently. Though the FTP site provides the README file, it can't keep a current step in updating the information to the parallel files and directories.

### **3.7 Inconsistency in naming customs**

No matter from the Archie's retrieval results or the Project Gutenberg documents, we can discover that there's no uniformed naming rule to the e-documents universal on Internet. The same problem also occurs to the Gopher and WWW space. Under such as situation, the IETF (Internet Engineering Task Force) has begun to contrive the URN (uniform resource name) to solve this problem. Meanwhile the libraries also devise the Metadata scheme to offer a solution to this same problem.

### **3.8 Difficulty in judging from the printings to e-documents**

In Project Gutenberg, all of e-documents with an announcement tell user the original printings what the e-document based on to produce electronic versions. When the netusers got these e-documents, they would make a clear distinction between printings and edocuments. But not all e-document are with the same or similar announcement as the above mention on Internet.

### **4.0 Standards and derived problems in practice**

When librarian uses the AACRII Rules and MARC Formats to organize the Internet resources, several existed standards can be taken. Could be no problem when the libraries apply these standards to cataloging and classifying electronic documents? Certainly not. Some problems come from standards, some are from data. Let's review them by the seven key points:

#### **4.1 Difficulty in identifying different versions of e-documents**

The electronic or digital information has one common distinction, that is editable with the appropriate editing software. Furthermore, the electronic one can be delivered to anywhere instantly by network's connections. It has become a big problem than ever before to distinguishing the different versions of e-documents whether is same or not and brought out many great troubles to the cataloging processes. One of the apparent examples is how to judge the difference between electronic and printing, or between ASCII and Postscript file formats that information they carry and represent.

#### **4.2 Data integrity**

It is almost impossible to determine whether the information is true and unchanged in electronic format, because it is editable with the right computer software. This problem caused the library have to spend much time, more jobs and requirements in judging information identity. If ignoring that, the catalog should be an error introductory guiding user to the wrong information and cause to make an incorrect conclusion in research.

#### **4.3 Different solutions to read out the diverse file formats**

The first requisite step to organize the electronic documents is to read out the file format

accurately and then get the title screen to catalog. There're many existent file formats achieved to different requirements on Internet. Of course this situation brings out many kinds of resolutions to read file, so the staff who is charging this job has to learn file resolutions well. Otherwise, no document can be done to be part of catalog.

#### **4.4 Instant change to information**

It is a fact to believe that the Internet has also changed the pattern of scholarly communication model. 'Here one second, gone the next.' and 'content is updated frequently' are two specific characteristics to the electronic documents, especially saved and accessed on Internet. Hence, there're two grand challenges to the catalog in updating the classification number, subject heading and location (i.e. URL) to be current, accurate of information content and position. It will be a continuous process to maintain these items for cataloging. If the catalog is unable to achieve currentness and accuracy, the users has to face two problems, that is, can't get and access to the correct information instantly.

#### **4.5 Diversity both in amount and dimension to networked documents**

It is true that no one can offer an exact statistical data telling how many resources exist on Internet. It is impossible for anyone or unit to organize them independently as well as the library. As for this, Vianne Sha presents four viewpoints worthy referring:

- \*organize and catalog Internet resources owned and maintained by the local system.
- \*organize and catalog significant research materials that have strong interests for the local patrons.
- \*organize and catalog significant tools that can improve reference services.
- \*organize and catalog significant tools that can update the knowledge and improve the skills of the library staff<sup>7</sup>.

In addition to these, the library also has to catalog the unique resource. If not does in this way, some special information will be lost quietly and users have to encounter information explosion as well as loss at the same time. Certainly it should be a gap to the cultural heritage.

#### **4.6 Ambiguity to definition between data file and program file in AACRII**

According to the chapter 9 in AACRII, there's no clear definition between the data and program file. The rule only gives us so indirect hint as Rule 9.0A1. Depended on a blur-ring line between data and program file, how can we ask cataloging staff describe them correctly and guide user to get information? If catalog wants to be so, we've to give clearer definition to diminish this ambiguity.

#### **4.7 Inadequacy to the Note Entry in AACRII rules**

If the cataloging staff follows the AACRII Rule's direction, and then the order of note entries are arranged and described as rule's listing (Rule 9.7B). It is unreasonable to ignore the e-document's characteristics and applications. For example, the System Requirements and File

---

<sup>7</sup> Vianne Sha, " Guidelines for Cataloging Internet Resources " ",  
<http://www.nlc-bnc.ca/documents/libraries/cataloging/sha1.Txt>

Characteristics are much more important than other kinds of note entries obviously. Therefore, this rule is definitely inadequate to the e-documents and needs to be revised.

## **5.0 Suggestion**

In this section I will present some suggestion to the AACRII rules and UFBD (USMARC Format for Bibliographic Data) after discussing the changes to the bibliographic control environment, analysis of the electronic Chinese documents and derived problems in practice. Furthermore, I will offer a model with necessary requisite fields to the networked resources on Internet for the bibliographic control.

### **5.1 Making a clear definition between resource and service**

The 'resource' and 'service' terms are always mentioned in UFBD's proposals and discussion papers, but have never been defined clearly to each other. What is the main difference between resource and service? Based on personal understanding, I present my opinions as below:

resource : is an object, and has various facets. It can be accessed by email, file transfer, or telnet, etc. An object may be a file, system, database, or server. So it is cataloged into the range from Oxx to 8xx blocks in UFBD, not including the tag856.

reservice : is a means to access the above object, for example email, file transfer or telnet. So it is cataloged into be the tag856 only.

From UFBD Structure, we can make a conclusion that resource is belonging to the bibliographic record level, so as service is belonging to the item record level. Therefore, resource has more than one service.

### **5.2 Drawing a bright line among author, producer and distributor**

On Internet it seems to be blurring among author, editor or producer and distributor within the new scholarly communication model. I try to offer one further definition to them as described as below.

author : is a personal or organization who is responsible to CREATE the intellectual works or documents.

Producer : is a personal or organization who is in charge of TRANSFORMING author's works into electronic form.

distributor : is a personal or organization offering the DEVICE or UTILITY to DISSEMINATE the author's works on network.

In UFBD the author is entered into lxx and 7xx, but the producer and distributor aren't put to be same tag. The producer should be cataloged into the \$f (manufacture) of tag 260, and the \$a (hostname), \$b (IP address) and \$n ( name of location of host on \$ x) in tag856 are for the distributor.

### **5.3 Changing the GMD terms**

No matter in AACRII or UFBD, the 'computer file' is both for local and remote electronic files in computer. In essence, all of them are materials in an electronic format. If we change the term from 'computer file' to 'electronic resource' and that may be more appropriate for local's computer file and Internet resource.

### **5.4 Offering clear explanatory rules among Material Specific Details Area, Physical Description Area and File Characteristics of Note Area**

These three are so close and dependent on each other and have a hierarchical relationship among them from AACRII rules. Taking a look between \$f(Filename), \$s(File Size) of tag 856 and Physical Description Area, we'll discover they are duplicate. It can be considered to replace the Physical Description Area with \$f and \$s in tag856 during cataloging Internet resource. Secondly, it is impossible that we use these terms such as 'computer data' in the Physical Description Area by taking the suggesting from OCLC's Assessing Information on the Internet Project <sup>8</sup>, because the Project doesn't give us the clear definition to these terms. Moreover, the speed of Internet resource to change is so rapidly. It seems not to be workable for such suggestion. Perhaps it could be entered by any current terms to reflect the rapid change of Internet resource.

### **5.5 Cancelling the order in describing note**

According to the AACRII rules, we have to follow the rule's directions to record the different notes orderly. But these restrictions are unreasonable and unpracticable. Because the different notes have various functions and effects on the data description. For example, The System Requirements and Restrictions Notes are much more important than others. My suggestion is to describe them depended on library's request and data characteristics, not same as the rules.

### **5.6 Careful use between \$n (Distribution of Resource Service) of tag856 and Country Code of tag008**

These two codes seem to be identical and duplicate, but not in fact. The country code is for the resource in bibliographic record level, but the \$n of tag856 is for the service in item record level. These two are radically different in identity and we can't confuse them together in any way.

### **5.7 Diminishing the duplicate function between \$f (Electronic Name) and \$g (Electronic Name - End of Range) in tag856**

These two subfields are to record the resource's file name. If the amount of file is more than two, then we use \$g to describe the last file name to indicate the file range and use \$f to describe the first file name. If the file is only one, we just use \$f to record it. The above two ways are needed to be discussed furthering, because it is unreasonable to use two subfields to distinguish the same situation between one and more than one file. It seems to be practicable in using only one subfield to accommodate this situation.

---

<sup>8</sup> *Martin Dilon, et al., Assessing Information on the Internet : toward providing library services for compute r-mediated communication (Dublin, Ohio : OCLC, 1993 ), pp. 5-6*  
*FTP : ftpftp.rsch.oclc.org at/pub/internet-resource-project/proposal directory*

### 5.8 Deliberation in using two tag008s

It is a future tendency to use two tag008s indicating the data characteristics, and Priscilla Caplan<sup>9</sup> also claims the similar one. According to my investigation, I think we've to use two different tag008s to two kinds of materials, not to same one. For example, the map book is with the characteristic both of map and book and we have to use two tag008s(map and book) to describe it. Especially to the Internet resource, many are reproduced from printing to electronic format.

### 5.9 Redefining the content of tag008

The Assessing Information on the Internet Project has presented its proposal to revise the tag008 content, but it is needed to practice for a long time to finish, test and reveal its appropriateness. While revising the UFBD, it should be revised thoroughly to indicate the right and multiple content of documents characteristics, not just for the Internet resource.

### 5.10 Defining a prototype for e-documents bibliographic information

Based on my survey to the e-documents of the Project Gutenberg and TANet's FTP sites, some traditional bibliographic items should be presented at document, for example author, publisher, etc., but some are new. I offer one model with requisite elements and an example for cataloging the e-document as below:

title	25 dynastic histories
author	
publisher	Institute of History and Philology, Academia Sinica
producer	Computing Center, Academia Sinica
distributor	Academia Sinica
edition	Taipei, Ting-wen
original source	25 dynastic histories
file size	39,969,533 Chinese Characters
file format	ASCII files with SGML markup
file resolution	Chinese Text Retrieval System
character set	Big-5 Chinese Character Set
keyword or subject-heading	History
abstract	
note	WWW user interface
access method - URL, 856	hppt://www.sinica.edu.tw/fms-bin/ftmsw3
TOC	

### 6.0 Noticeable trends

Through the library approach to cataloging Internet resource has various characteristics, there're new trends to be alert to improve the catalog. They are outlined as the following points:

---

9 Priscilla Caplan, " USMARC Format Integration, Part I : what, whY and when?" *The Public-access Computer System Review* 3(5), 1992:34 Email GET CAPLAN PRV3N5 LISTSTERV@UHUPVMI.UH.EPU

### **6.1 Automatic gathering and updating information**

By application of WWW, so many existed WWW search engines help user mining out information from Internet and these engines are called spider, robot, crawler, etc. Besides the search engines have an ability to gather information automatically, the results that they offered are current. The Z39.50 application also can achieve to the same purpose. In other words, these two applications have some specific points as below and the current catalog has to comply with :

- automatic gather information
- automatic update and offer current status information
- access to information virtually by many and saved only one copy physically.

### **6.2 Image retrieval**

Traditionally we catalog the image data by appended with the bibliographic data, and then users only search to the partial facet of image data. Now the information retrieval technology is improved and can retrieve out more details of the image shape, color, texture, position, etc. The Query by Image Content (QBIC) that IBM invented is one of top image retrieval technologies around the world. Certainly the image retrieval has a heavy impact on catalog in the future.

### **6.3 Metadata - Dublin Core**

In March 1995, the OCLC and NCSA co-sponsored the Metadata Workshop and convened 52 selected researchers and professionals from librarianship, computer science, text encoding, and related areas to develop the resource description records for networked electronic information objects<sup>10</sup>. Why is it so important? Because it has five functions will bring impact on the cataloging. They are:

- encouraging authors and publishers to provide metadata and that can be collected by automated resource discovery tools
- creating a template with metadata elements for network publishing tools
- serving as the basis for a more detailed cataloging record
- being understood across user communities if metadata become a standard<sup>11</sup>
- becoming a similar MARC device among HTML, SGML, TEI Header, MARC and URN (uniform resource name), etc.

### **7.0 Conclusion**

It is out of the question that Internet has become another information treasure and founded the base of the digital library. If without the appropriate bibliographic control, the users will face information explosion and loss simultaneously in the near future. The librarianship also has the related practiced experience and foundation of organizing diverse data, especially in cataloging and classification field. It becomes our obligation to organize the Internet resources effectively and efficiently.

---

10 Stuart Weibel, et al., " OCLC/NCSA Metadata Workshop Report, "

[http : //www.oclc.org:5046/conferences/metadaldublin\\_core\\_report.html](http://www.oclc.org:5046/conferences/metadaldublin_core_report.html)

11 Ibid