Creating Web Pages in Japanese

Eddy Harrison

Creating web pages that contain Japanese text can present some challenges in the areas of character input and character code compatibility. With older operating systems such as Windows 95, 98 and NT one needs to be aware of issues such as language encoding (shift JIS, for example) and fonts that can display Japanese properly. Fortunately, Windows 2000 is fully Unicode compliant and web authors will not have to face the assortment of problems associated with the earlier operating systems.

The Windows 2000 and XP Environment

As long as one has access to Japanese friendly web editing software that works with Windows 2000 and above, the job of preparing a web page with Japanese language elements will not be too difficult. In fact, it is possible to create a variety of pages without writing any HTML code at all.

Two things, however, do need to be understood: how to use the web editor, and the process by which html files are transferred to the server that contains your web pages. It is *essential* to gain a solid knowledge of these two areas before creating your web pages.

The easiest and most reliable way to create a web page that contains Japanese is with Front Page 2002 running under Windows 2000 or Windows XP. Web pages created in Front Page with the MS Ariel Unicode font will result in a Unicode page viewable with little or no difficulty on most browsers. The page can be created in a WYSIWYG (what you see is what you get) editing environment without the author having to write any HTML code or worry about language encoding. I strongly recommend that you do your web work with the tools mentioned above. They will save you hours of work and frustration.

Having said that you can create a web page without ever writing a line of HTML code, I should comment on the other side of this coin. While it is perfectly true that simple and complex web pages can be created solely with a web editor, some people feel that a knowledge of HTML is nonetheless important and even necessary. Without HTML skills you are limited to what the web editor produces for you. You will not be able to make any adjustments on your own or recognize and repair any problems that can't be addressed through the web editor.

Let us say that you are creating a page with text, tables, and graphics and the web editor does not produce the visual result you expect – no matter how many times you re-work the input. For example, a cell in a table might not expand to the exact size you need, or a photograph or block of text may center just a little too high.

Likewise, a conflict with style sheets may be altering the appearance of your document Most often, these kind of problems can be solved by making adjustments directly to the HTML coding.

Whether or not it is worth the time and effort to learn HTML just to make a periodic adjustment is a decision you will have to make. However, if you plan to do a lot of web work on your own, you will be well served with at least a basic knowledge of HTML.

Creating Web Pages

Once you have learned to use Front Page or a similar web editor, you should be well on your way to making Japanese Web pages. It is essentially a matter of creating the pages in the web editor and uploading them to the server.

A word of advice about Word. While it is true that Microsoft Word has the ability to create web documents in English and Japanese, there are two reasons to avoid this option. If you allow Word to save a document as HTML, it will convert your text into a *very large* XML/HTML hybrid web document. The coding is quite complex and truly voluminous. Many people refer to this as HTML bloat. A file created in this manner takes a lot of unnecessary space on the server and will not please most system administrators. The other problem with Word generated coding is that it is so complicated that it may be extremely difficult or even impossible for you or others to make manual adjustments in the document without an advanced knowledge of HTML and/or XML.

The degree of HTML bloat can be lessened in the Office XP version of Word by saving the document as a filtered web page. The resulting code is much more compact and quite acceptable for most documents with Front Page or a document that was hand coded in HTML, make some changes, and allow Word to save the document as HTML, the original coding will be greatly altered – HTML bloat. You may experience some unexpected and unwanted changes in the appearance of your document that will be very time consuming to restore.

To avoid this, make the changes in your web editor or, if you are doing them manually, use Notepad to make and save the changes.

Windows 98 and NT Environment

If your library does not support Windows 2000 or above and you must work in an earlier operating system you can still create good web pages with Japanese. You can use a web editor such as Namo which works with Windows 98 and NT and can accommodate Japanese. The process and resulting web page will be very similar to those created with Front Page in Windows 2000.

If you have access to Office 2000 on your Windows 98 or NT machine, it is possible to create a document in Word using the Global IME to input the Japanese. In this

case, it important to save the document as “encoded text” and change the file extension later to “.html”. This will prevent Word from converting the text to its own version of HTML bloat. The resulting web page will be a mixed text, Unicode page viewable on most browsers. Adding or modifying Japanese text should be done by pasting from Word into Notepad to avoid the HTML bloat that occurs when you edit and save directly in Word.

Documents created with Japanese Windows 95 /98 or the Japanese IME for the English version of Word 95/98 will most likely be in Shift JIS encoding. In most cases, these documents must be edited with a Shift JIS input device. Mixing Japanese Unicode text generated by Word 2000 in a Shift JIS document will cause display problems.

CD-ROMs

As a rule, libraries usually prefer databases available on the web rather than CD-ROMs because they do not present physical access problems and are comparatively easy to administer. The web is well suited to such large collections as newspaper back files, full text archives, and regularly updated indexes. Web access has many good points, however, for the near future, CD-ROMs and, eventually, data DVDs in one form or another, will continue to be found on the market. Publishers in Japan, and other countries for that matter, like to issue their products on CD-ROM for a variety of reasons.

A CD-ROM product involves a one-time production expense and very little on-going costs. Once the CD-ROM is sold, payment is collected and that is basically the end of the transaction. Technical support is not often needed or offered after the initial installation. In addition, some publishers in Japan feel that data provided on a CD-ROM is more secure than on the “free-wheeling” web where it can be easily reproduced and manipulated by anyone with the skills to do so.

There are many expenses associated with a web-based product. Site hosting and maintenance, customer accounts, billing activities, and support services are among the on-going costs. These expenses don’t go away after the product is released. Both the provider and the end user face and continual investment as long as the database is available on the web though a web product might be more convenient to access and use, some vendors will still prefer CD-ROM as a delivery medium.

For those who find this difficult to believe, consider the fact that the Yomiuri Newspaper decided to offer the Meiji and Taisho back files on CD-ROM, rather than on an integrated web site. The same is true for the Vatican newspaper, *L'osservatore Romano,* which was issued in 141 CD-ROMs.

Access Methods for Japanese CD-ROMs

Most Japanese CD-ROMs will work on a PC running under Windows 2000 or above. The default language must be set to Japanese for proper display of characters in menus and most portions of the text. This is an important requirement to

remember. English language CD-ROMs will run with little or no problem with the language default set to Japanese. However, the reverse is not true.

Options for public access on PC workstations are:

Insert CD-ROM for Each Use:
Many CD-ROMs have an auto run feature and users can simply insert the disc in the drive and use it without going through an installation routine. On the other hand, some CD-ROMs require the user to make basic installation decisions each time the disc is run. It is not convenient or desirable to have patrons installing start programs on PCs belonging to the library (although it is done at some institutions). It is usually possible to copy the start program onto the hard disc and create a shortcut that prompts the user to insert the disc in the CD-ROM drive. This enables the patron to use the CD-ROM with only two steps; clicking on the icon, and inserting the CD-ROM. From a public service and security standpoint this is preferable to allowing patrons to "install" even a minor start program with each use.

Copy Entire CD-ROM onto Hard Drive
In some cases it is possible to copy the entire CD-ROM onto the hard drive, although this invariably requires some software "tweaks," especially in the start program. Once installed successfully, this method insures easy access and avoids handling of discs by patrons and staff. A small number of applications require the physical presence of the CD-ROM disc in the drive in order to run. Although this can sometimes be overcome, it can be difficult and time consuming. In a few cases that I am aware of, it has proven impossible to make adjustments to enable direct access from the hard drive.

Mount CD-ROMs on a Server
Most Japanese CD-ROMs can be run from a server with Windows 2000 or XP, provided the language default is set to Japanese. If your library has a central server this might be a problem since the system administrator will probably not be eager to set the default to Japanese. If you are fortunate enough to have the financial and technical resources, a dedicated server for Japanese CD-ROMs can be installed in the library. This, of course, requires on-going technical support and maintenance.

Preservation Issues
To some degree, the jury is still out on long term durability of CD-ROMs, but high quality products have an estimated life expectancy of 100 years. Coating technology used on CD-ROMs has improved greatly in recent years, but the discs are still subject to varying degrees of wear. Aside from possible surface deterioration due to oxidation, the major concern is handling. Obviously, discs should be stored in jewel cases and kept away from the usual environmental hazards. It is most important to prevent scratching and acid deposits from fingers on the reflective (the "mirror-like surface") of the disc. The label surface of the disc is comparatively thin and scratches or gouges can damage the data layer making the disc unusable.
CD-ROMs will eventually be replaced by data DVDs, however the good news is that the DVD players are expected to be able to run the older CD-ROMs.

DVDs

DVDs (known also as digital versatile dis**cs** or digital video dis**cs**) are very similar to CDs, but have a much greater storage capacity. A standard DVD can contain about seven times more data than a CD-ROM. The audio and visual quality of a DVD is far superior to that of a video tape cassette and there is no doubt that DVD will gradually replace the cassette just as CDs replaced vinyl LP records. In fact, the transition has already started for motion pictures.

This would normally not be a matter of concern for Japanese studies librarians except for the regional restrictions that motion picture producers and distributors have managed to impose on the manufacture of DVDs and players worldwide. The world has been divided into six regions for the purported purpose of controlling the release date of motion pictures in theaters and on DVDs.

A region locking code is included in all standard DVD players and most DVDs also contain a region specific code. The codes much match in order for the DVD to play. If the codes do not match, the DVD will not be viewable.

For example, the United States is in region 1 and Japan is in region 2. Even though both countries use the same NTSC video format, the regional code will prevent a region 2 Japanese DVD from being played on a region 1 DVD player in the United States. The reverse is also true.

The solution to this problem is to acquire a code free DVD player in which the coding device has been disabled. These players are not illegal in the United States, however they are not available in conventional retail outlets. They can be purchased online and at some small specialty stores. With a code free DVD player, it is possible to watch a movie on DVD from any region in the world.

There are several implications here for Japanese collections outside of region 2. The most obvious is that the majority of library patrons will not have a region free DVD player and will not be able to view the movies on their home or workplace equipment. To accommodate Japanese DVDs, the library would have to provide a region free DVD player for on-site use and viewing would be restricted to a specific machine. Likewise, a cinema studies instructor would have to make sure that his/her classroom could be equipped with a region free DVD player in order to show a Japanese film to a class. These restrictions will surely reduce the utility of the library's Japanese film collection. One could attempt to make a video tape copy for instructional purposes, however aside from copyright issues, an anti copying mechanism called macrovision may well cause the VHS copy to be distorted. There are ways around this, but again, a region free DVD player or a device to over-ride copy protection is necessary. In either case, concerns about copyright are still very much present.
Librarians will be wary of building a collection of Japanese DVDs with regional codes imbedded in them. In the long term, regional locking of DVDs may not be a commercially sustainable marketing system and could disappear. If this were to

happen, new machines may or may not be able to play the "old" region locked DVDs. That would mean that the library's collection of region 2 Japanese DVDs would not be playable on a growing number of "new" machines and could eventually become obsolete if not transferred to another format. On the other hand, the "new" players might just play DVDs from all regions. This would make sense, but there is no clear idea of what the future holds in this regard.

SGML, XML, TEI, EAD
A markup language alphabet soup

Acronyms expanded

- SGML – Standard Generalized Markup Language
- XML – eXtensible Markup Language
- DTD – Document Type Definition
- HTML – Hyper Text Markup Language
- XSL – eXtensible Stylesheet Language
- EAD – Encoded Archival Description
- TEI – Text Encoding Initiative
- ASCII – American Standard Code for Information Interchange

What are markup languages?

- Codified conventions for describing attributes of a text within the text itself.

Visual vs. content markup

- Visual markup relies on humans to interpret, since the display only implies structure and content.
- Markup languages like SGML and XML aim to describe the structure and content of a text in explicit terms, so that software systems can interpret and manipulate the text without ambiguity.
- Stylesheets are then used to apply the visual markup to the text based on the structural and content markup.

What are SGML and XML?

- Not markup languages in themselves, but systems within which to define markup languages.
- A set of principles, syntactic rules, and components with which to build markup languages.

Components of an SGML or XML system

- Document Type Definition (DTD) *or* schema – *the grammar rules*
- Application guidelines – *the stylistic guide*
- *Document instance – the SGML or XML file itself*

- Elements – *the information between particular* tags
- Entities – *info pulled from outside the* instance
- Attributes – *characteristics of an* element
- Stylesheets – describe how to display the instance

Beginnings of SGML and Early adoption of SGML

- 1969 –Charles **G**oldfarb, Edward **M**osher, and Raymond **L**orie develop GML at IBM to allow different systems to share documents.
- GML is based on ASCII and meant to be platform independent -- readable by any system -- and understandable by humans. Longevity and portability of data and metadata are goals.
- Introduced concepts of formally defined document types with explicit nested element structures.
- Invented the <tag>syntax</tag>
- 1978 - Goldfarb transforms GML into a broader standard, publishing first draft of SGML in 1980.
- 1986 - SGML accepted as an international standard (ISO 8879:1986)

SGML reaches academia – TEI

- 1987 – TEI (Text Encoding Initiative) begins to define a system for encoding texts for use in academic study.
- 1990 - First draft of TEI standard published.

SGML reaches archives - EAD

- 1993 – Daniel Pitti at UC Berkeley begins work on FINDAID DTD to provide a mechanism to standardize encoding of archival description (finding aids).
- 1998 – EAD Version 1.0 released as an XML DTD.
- 2002 – Hundreds (thousands?) of archival institutions now use EAD as their main descriptive apparatus

SGML, HTML, and XML

- SGML still not very widely adopted because of complexity (**S**ounds **G**reat, **M**aybe **L**ater).
- HTML (Hyper Text Markup Language, a very basic SGML DTD) takes off, but goes against many of the principles of SGML, since its tags are mainly visual.
- SGML community develops XML (eXtensible Markup Language), a simplified version of SGML, to help increase adoption for semantic and structural markup of networked resources.
- XML becomes the “next big thing” both for data storage and for data interchange, with many other technologies growing around it (XSL, XSLT, FOP, SOAP, etc.).

SGML / XML and Libraries
Why are SGML and XML important for libraries?

- Metadata based – encourage thinking about and describing content first, display later
- Open standards – not hidden and not owned by any organization
- Platform independent – can be used on any computer system
- Human-readable – you don't need any special tool to interpret
- Flexible - easily transported and converted to other formats

For further information, Google these terms
On XML and SGML

- Gentle Introduction to SGML
- XML Cover pages
- XML FAQ
- XML MARC
- EAD
- TEI

Other good sites on XML and their use in libraries are:

http://www.slis.ualberta.ca/538-99/cbradley/xml.htm
http://xmlmarc.stanford.edu/LJ/
http://xmlmarc.stanford.edu/
http://www.albany.edu/~gilmr/
http://www.albany.edu/~gilmr/metadata/