

DRAFT

UNC Digital Library Services

Digitization Guidelines [Full Version]

Introduction2
 Using these Guidelines.....2
Before You Begin Digitizing3
Images4
Photographs, Paintings, Slides : Raster Images.....4
 Images for Archiving4
 Images for Web Delivery.....4
 Images for Printing.....5
Logos, Line Drawings, Geometric Shapes : Vector Images.....6
Audio.....7
Music Audio7
 Archiving7
 Web Delivery7
Voice Audio8
 Archiving8
 Web Delivery8
Video9
 Web Delivery and DVD9
 Video for archiving10
Text.....11
 Web Display.....16
 Printing.....16
 Archiving.....16
Presentations18
Other19
Storage20
Looking for Something Else?21
Campus Help **Error! Bookmark not defined.**

Introduction

The purpose of this document is to help support common digitization needs at UNC-Chapel Hill. The recommendations included here may not be appropriate for all digitization projects, especially those with special needs. The recommendations are likely to evolve with advances in digitization techniques and underlying technologies.

These guidelines are part of an effort to encourage campus digitization projects to adopt standards that will facilitate resource-sharing within and across departments. A central infrastructure for storing, managing and sharing media collections is currently being piloted on campus. For more information, see the Storage section of the document or visit the UNC Digital Library Project website at <http://www.unc.edu/projects/diglib/>.

The recommendations in this document were developed by a working group of media users and campus support representatives. The following organizations participated, and are adhering to these guidelines in their support of departments and individual instructors and researchers:

- Academic Affairs Library
- Center for Instructional Technology
- Center for Teaching and Learning
- College of Arts and Sciences (OASIS)
- Health Sciences Library.

Using these Guidelines

This is the full version of the UNC Digital Library Services' Digitization Guidelines. The Quick Reference version provides the standards for each media type along with brief explanations. This version provides richer background information and explanations, and explores a number of emerging digitization solutions not covered in the Quick Reference version.

Before You Begin Digitizing

Before you digitize anything, take some time to consider your needs. The worst possible outcome is to spend time digitizing materials that end up being inappropriate for the goals of your project. To avoid this scenario, consider a number of issues ahead of time.

- For what purposes will the materials be used?
- What level of media quality is necessary to achieve your goals for the project?
- Who needs to have access to your digital media?
- What options do you have for making the materials available to your audience?
- Who owns the copyright to the materials you are digitizing?
- What options, both short-term and long-term, are available to you for storing your digitized media files?
- Would it be cost-effective to outsource the digitization of your media?

If you have any questions or would like to discuss a digitization project with someone on campus, contact one of the organizations listed under Campus Help or send email to mediasupport@unc.edu.

For additional information on planning digitization projects, you might also consult one of many online resources available. Several are listed here:

- Digital Library of Georgia Digitization Guide (<http://dlg.galileo.usg.edu/guide.html>)
- Washington State Library's Digital Best Practices (<http://digitalwa.statelib.wa.gov/newsite/projectmgmt/planning.htm>)
- Digitization Guidelines at Harvard University (<http://preserve.harvard.edu/resources/digital.html>)
- National Digital Library Program Checklist (<http://memory.loc.gov/ammem/techdocs/prjplan.html>)

Images

In computer terminology, images are referred to as either raster-based or vector-based.

Raster images, sometimes called bitmap images, are usually used for photographs, paintings, slides and other images that have subtle gradations of color. Raster images are good at preserving nuances of shading in images because they see images as grids of pixels (picture elements, the smallest units of an image). The drawback to them is that resizing raster images, especially enlarging them, produces a loss of image quality.

Vector are usually used for things like logos or simple line drawings. Unlike raster images, vector images maintain their quality no matter what their size. You can make vector images larger or smaller at will. The limitation of vector images is that they do not allow for the subtle shading that we usually associate with photographs.

Photographs, Paintings, Slides : Raster Images

Images for Archiving

When you enlarge digital versions of complex and colorful images like photographs, paintings and slides, the quality of the scan degrades. Having a larger archive version will provide you with the most flexibility when you need to go back and create new versions of an image.

The major consideration in archiving images is that the images not be stored in a *lossy* format - one that compresses the image data in order to save space. JPEG is *lossy*. It retains a decent enough quality image for the purposes of web viewing, but the image contains distortions necessary to shrink the file to a more web-deliverable size. The preferred cross-platform format for *lossless* images is the TIFF. However, the TIFF stores color data *for each pixel* in the image; this means that the file can become large very quickly. Before attempting to archive a TIFF image, it is important to have a storage facility that can accommodate your project. As a benchmark, a 1800 x 1200 pixel TIFF at 16 million color depth will average between 3 and 10 MB. While Photoshop's PSD format is widely used for editing, it is proprietary and thus not recommended for archive use.

Images for archiving

Format	TIFF (no compression), PNG
Size	4000x2500 to 6000x4000 pixels
Campus Help	CIT, CTL, OASIS, UNC Libraries
Recommended Editors	Adobe Photoshop, Adobe Photoshop Elements, Microsoft Photo Editor
Recommended Viewers	any web browser

Images for Web Delivery

For images that you plan to make available via the Web, it is recommended that you save these files in the JPEG format. JPEG is the ubiquitously accepted file format for display in current web browsers and operating systems, it is capable of displaying the millions of colors necessary for rendering photographic images properly, and file size is relatively small due to its built-in compression.

DRAFT

Ideally, your JPEG images would be derivatives of your archive version, or original scan (see Images for Archiving). You should also have two versions of each image, one larger than the other. There are several advantages of having two different versions of your images for use on the web. Smaller versions are easier for users with slower speed Internet connections to download. The larger version allows for examining details of the image.

If you are affiliated with the College of Arts and Sciences and are having media scanned through OASIS Media Services, your web-ready versions will be created for you. If not, some scanning software and common image-editing software like Photoshop can be used to create these derivatives from your archive version. For assistance, contact mediasupport@unc.edu.

Images for web delivery (normal)

Format	JPEG
Size	600x400 pixels
Campus Help	CIT, CTL, OASIS, UNC Libraries
Recommended Editors	Adobe ImageReady, Adobe Photoshop, Adobe Photoshop Elements, Macromedia Fireworks
Recommended Viewers	Microsoft Photo Editor any web browser

Images for web delivery (large)

Format	JPEG
Size	1800x1200 pixels
Campus Help	CIT, CTL, OASIS, UNC Libraries
Recommended Editors	Adobe ImageReady, Adobe Photoshop, Adobe Photoshop Elements, Macromedia Fireworks
Recommended Viewers	Microsoft Photo Editor any web browser

Images for Printing

If you are printing images for your personal use, or use in a class, we recommend that you follow the Web Delivery guidelines (above).

If you are readying images for publication in a journal, book, magazine, etc., publishers generally have much higher standards for image quality. The standard size for formal publications usually varies anywhere between 150 dpi and 300 dpi. You should contact your publisher and inquire about its preferred specifications.

Additional information regarding dpi or ppi and image size:

Image resolution is a tricky and often confusing concept. Typically, image resolution is defined by dots or pixels per inch (dpi or ppi respectively). However, this terminology is only useful for determining the *quality* of an image when it is *printed*. The more dots or pixels per inch, the more complex and detailed the image will appear *in print*. But if you are simply displaying an image on a *computer screen*, the resolution is only referring to the *size* of the image on the computer screen. When you change the resolution of a digital image, it does not become more or less complex - it only gets bigger or smaller. The standard 300 dpi for formal publications ensures a level of quality for the printed image, regardless of its eventual size on the printed page. Below you will find a table

DRAFT

of inch-sizes, dpi, and resulting pixel length and width, and a table with recommendations for some common photographic images.

Original	DPI	Result
Slide (1" x 1.5")	1200	1200 x 1800 pixels
Print film (4" x 6")	600	2400 x 3600 pixels
Oversized print (8" x 10")	300	2400 x 3000 pixels

Images for printing

Format	JPEG; TIFF; PSD
Size	300 dpi
Campus Help	CIT, CTL, OASIS, UNC Libraries
Recommended Editors	Adobe Photoshop, Adobe Photoshop Elements, Microsoft Photo Editor

Logos, Line Drawings, Geometric Shapes : Vector Images

Unlike raster-based images, true vector-based images do not require different sizes for different output formats. Geometric shapes are the basis of vector-based images, and as a result, it is possible to enlarge true vector-based images to any size without loss of quality. The most common format for vector-based images today is the CompuServe GIF. However, the CompuServe GIF is not technically vector-based (it is, rather, a raster/pixel-based format). It does not provide the same quality and malleability as true vector-based formats.

There are emerging options for working with vector graphics in a web environment, but each requires a browser plug-in to display the image properly. This is likely to change in the future, but right now the following is recommended: If you need to display a logo or simple line drawing on the web today, the best format is GIF (although it is not as scaleable as Flash or SVG). If you have some image and software experience, and your students (or viewers) do not mind downloading the Flash plug-in, then Flash is a good option. If your project is not taking place for a few years, or you want to be assured of future compatibility, SVG is your best choice.

Vector-based Imagery

Format	GIF*, Flash, SVG
Size	N/A
Campus Help	CIT, CTL, OASIS, UNC Libraries
Recommended Editors	Adobe Illustrator, Macromedia Flash/Fireworks
Requirements	Any web browser (with Adobe SVG or Macromedia Flash plug-in)

Audio

When dealing with audio, it is important to make some decisions early on regarding what exactly the audio stream will contain, and who will be hearing it. Your decisions will be different depending on the content of your audio stream. For example, the human voice (giving a lecture) has different requirements than most music. The three most important factors to remember when making your decisions are: bit-rate (defined by your choice of format), frequency (KHz, which is format-independent), and bits-per-channel (mono vs. stereo). The higher these are, the more sonic range the audio file can contain (and the more space it will use). Basic audio files, including AIFF and WAV do not allow for variable bit rates, and by default these formats contain the highest bit-rate possible. On the other hand, newer compression-friendly formats like MP3 allow for variable bit-rates, and therefore variable file sizes.

Music Audio

Archiving

You can not convert lower-quality digital audio back into a higher-quality version. Keeping a high quality archive version will provide you with the most flexibility when you need to go back and create new versions of a sound file. The base unit for archiving music remains the uncompressed WAV (for PCs) or AIFF (for MACs) file, at the highest frequency and bits-per-channel (i.e. mono vs. stereo) possible. Specifically, we recommend that audio be archived at 44.1 - 48 KHz with 16 to 24 bits per channel. If you have plenty of storage space, archiving these raw archival music files in a "native" format (WAV / AIFF) is optimal. However, if storage is an issue, there are emerging lossless file formats which decrease file sizes by 40 - 70 percent: shorten (SHN) and Monkey's Audio (APE) formats both decrease file size without the lossy compression inherent in MP3s.

Music Audio for archiving

Format	Windows Waveform (WAV); Macintosh AIFF; Shorten (SHN); Monkey's Audio (APE)
Size	44.1-48KHz; 16-24 bit; Stereo
Campus Help	CIT, CTL, OASIS, ibiblio.org, UNC Libraries
Recommended Editors	Goldwave or Soundforge (editing), RealMedia's production applications (encoding), Digidesign ProTools
Recommended Players	Any audio player

Web Delivery

If you are going to deliver music over the web, we recommend encoding it into MP3, or a comparable compressed format (like Real Media or Quick Time).

Music Audio for web delivery

Format	MP3, RealAudio, QuickTime
Size	128-224 Kbps / 22-44.1 KHz; Stereo
Campus Help	CIT, CTL, OASIS, UNC Libraries
Recommended Editors	Goldwave or Soundforge (editing), CoolEdit, RealMedia's production applications (encoding), Digidesign ProTools
Recommended Players	RealOne Player, WinAmp, Windows Media Player
Requirements	MP3 player

Voice Audio

Voice audio does not require the range or rate that music does. You can save both storage space and download time by compressing voice audio at lower frequencies (11 - 22 KHz). Single voice streams should also be recorded in monophonic rather than stereophonic sound, which immediately decreases the file size by half.

Archiving

We want to archive at the highest possible quality, but voice audio is not as demanding as music audio. Voice can be recorded in a monophonic sound, and at 8 bits per channel. We recommend archiving voice audio in the native format (WAV / AIFF), possibly burned directly to CDR as a good means of retaining the information. Note: Not all CDRs are created equal.

Voice Audio for archiving

Format	Windows Waveform (WAV); Macintosh AIFF; Shorten (SHN); Monkey's Audio (APE)
Size	22 KHz; 8 -16 bit; Stereo/Mono
Campus Help	CIT, CTL, OASIS, ibiblio.org, UNC Libraries
Recommended Editors	Goldwave or Soundforge (editing), RealMedia's production applications (encoding), Digidesign ProTools
Recommended Players	RealOne Player, WinAmp, Windows Media Player

Web Delivery

Voice Audio for web delivery

Format	MP3, RealAudio, QuickTime
Size	28 - 128 Kbps / 11 KHz
Campus Help	CIT, CTL, OASIS, UNC Libraries
Recommended Editors	Goldwave or Soundforge (editing), RealMedia's production applications (encoding), Digidesign ProTools
Recommended Players	RealOne Player, WinAmp, Windows Media Player
Requirements	MP3 player

Video

Video creation is one of the most complex multimedia activities you can undertake. The first things to determine, as with previous media types, are the input and output. What is the source of the video, and how do you want the viewers to see it? How much detail is necessary in the finished product? How long will this video project be in use? Consider some of the issues raised in *Before You Begin Digitizing* in advance of project implementation.

If you are creating video for long-term use, it is worth the time and effort to use the best possible video production equipment and exporting the highest quality video available. Currently, the DVD (or Digital Versatile Disk) is the highest possible quality. DVD can handle large video frames (720 x 480), and high quality audio (48KHz 24-bit). DVD's only drawback is that it uses major compression, so if you plan on re-using the video at a later date it is a good idea to retain an uncompressed copy of the original.

Web Delivery and DVD

There are many options available to those who want to deliver their video via the web. Depending on the ultimate audience, you could stream files using Quick Time (a once Macintosh-specific format, although becoming more common), Windows Media File (a Microsoft product), or RealMedia products. Because RealMedia files are small, retain excellent video information, and are platform independent, we recommend you use this format for web delivery of video, at a minimum resolution of 160 by 120 pixels. For higher resolution videos we recommend 320 by 240 pixels.

Optionally, for web delivery we also recommend a video standard comparable to VHS video: MPEG1-based Video CD (VCD), which is 320 by 240 pixels in NTSC format. The particular advantage of the VCD format is that it can be delivered to multiple media, including the television, with the relative ease of burning a CDR. While it doesn't stream as readily as RealMedia, it is comparable in video quality and compressed file size.

For higher speed networks, including intracampus as well as Internet2, we recommend MPEG2-based digital versatile disc (DVD) video. This video has a higher resolution (720 by 480) as well as better audio quality. (It is also our recommendation for archival video, so if the goal of the project is to produce video for use on DVD or over the internal campus network, choosing this format allows for both your delivery and archiving.

Video for web delivery

Format	MPEG1/VCD (DivX codec); RealMedia; QuickTime (Sorenson codec)
Size	160x120 to 320x240 pixels
Campus Help	ATN Video Services, CIT, OASIS, ibiblio.org, openVideo archive
Recommended Editors	Adobe Premiere, TMPGEnc, FinalCut Pro (Mac), RealMedia's production applications (encoding)
Simple Editors	Macintosh iMovie, Windows Movie Maker
Recommended Viewers	RealOne Player, Windows Media Player, QuickTime Player

Video for high-speed networks and DVD

Format	MPEG2/DVD
Size	320x240 (Web); 640x480 (VHS); 720x480 (DVD)
Campus Help	ATN Video Services, CIT,

DRAFT

Recommended Editors	OASIS, ibiblio.org, openVideo archive Adobe Premiere, TMPGEnc, FinalCut Pro (Mac), RealMedia's production applications (encoding)
Simple Editors	Macintosh iMovie, Windows Movie Maker
Recommended Viewers	RealOne Player, Windows Media Player, QuickTime Player
Requirements	greater bandwidth (e.g., Internet2); DVD-ROM

Video for archiving

Archiving raw video is very storage-intensive. Raw digital video AVI files require roughly 10 to 15 *gigabytes* of storage per hour! More reasonable is MPEG2-based DVD, which allows for a high quality video stream (720 by 480 pixels) as well as a high quality audio stream (384Kbps MP3 format). DVD-quality video is lossy and therefore not a traditional archival format, but its high quality makes it a reasonable alternative to the raw AVI format.

Video for archiving

Format	MPEG2/DVD; raw Microsoft AVI (DV-compatible); miniDV/DVcam
Size	720x480 pixels or higher
Campus Help	ATN Video Services, ATN Video Networking, CIT, OASIS, ibiblio.org, openVideo archive
Recommended Editors	Adobe Premiere, TMPGEnc, FinalCut Pro (Mac)
Simple Editors	Macintosh iMovie, Windows Movie Maker
Recommended Viewers	RealOne Player, Windows Media Player, QuickTime Player
Requirements	drive space (these files are very large)

Text

Many options exist for the creation and presentation of text. Text documents in their simplest form (plain ASCII, for example) have no formatting applied, but most of the commonly used document formats mix together the structural and presentation aspects of their documents. This is true of, e.g., Microsoft Word, WordPerfect, RTF, PDF, and HTML. There are also formats, particularly SGML and XML markup languages, which allow for the separation of these aspects. In the latter document types, the structure and semantics of the document are dealt with in the document itself, and the presentation is generally delegated to other software or documents. For the purposes of presentation, SGML and XML files may be transformed into any of the common formats mentioned above.

Anyone who wishes to produce a collection of texts should answer the following questions before deciding on what format(s) to use:

- 1) Who is the audience for this collection?
- 2) What is the level of complexity of the collection?
- 3) What text-creation capabilities do I possess or have access to?
- 4) How static are the documents in the collection and the collection itself?
- 5) How sustainable should the collection be?
- 6) To what kinds of uses might the collection be put in the future?

1) *Who is the audience for this collection?*

The answer to question #1 will determine what types of delivery mechanism(s) are needed and the format(s) in which texts should be delivered. In an instructional context, the most common methods of distribution are via the World Wide Web or print (in, e.g., a course pack or handout). If the texts are to be delivered over the web, HTML is an obvious choice. But there may be reasons why it is not a workable one. If the collection owner must be able to guarantee the same format for all users, HTML may not be viable, since the type of browser used to view it plays a large role in what the user sees. Likewise, if the documents must be printable, HTML is probably not the best choice. This is why online publishers typically provide both an HTML and a printer-friendly version (often PDF) of their documents. When this kind of multi-format publication¹ is necessary, most authors will want to create their documents in a single format and convert them to multiple other formats for delivery. Most word processing and desktop publishing software packages allow users to save files as HTML or to print to PDF. Likewise, documents produced in XML can be transformed into virtually any other format for publication.

Another question that follows from the first is: how is the audience expected to interact with the collection? If the collection is published on the web, the audience will need to have ways of finding particular documents or of finding relevant parts of large documents. At the very least, this will entail the creation of an index and/or table of contents. Some sort of searching feature may also be desirable. The size, complexity, and mutability of the collection will have an effect on the implementation of these ancillary tools.

2) *What is the level of complexity of the collection?*

Complexity may take a number of forms in collections of documents. The individual texts may themselves be complex documents. They may contain other types of media, such as images, for example, and they may have many internal and/or external cross-references, such as notes or

¹ [“Publication” in this context should be taken to mean any act of distributing documents to other people.](#)

DRAFT

critical apparatus. They may use multiple languages with different writing systems. They may be extremely large and/or have very complex structures. A collection may be complex because it contains many documents of different types with a web of relationships between documents. The level of complexity to be managed will shape the choices a collection owner makes about publication strategies.

size

For small projects without a great deal of complexity, standard word processing software (for example, MS Word, WordPerfect, and OpenOffice Writer) is likely to be quite adequate. All of these allow users to save as HTML or PDF for web publication. Very large (e.g. book-length) documents with more complex formatting requirements may cause ordinary word processors problems, however. For these types of projects, powerful desktop publishing packages like Adobe FrameMaker may be a better choice.

languages, encodings, and fonts

Different languages / writing systems are another source of potential problems and these become particularly acute across different computing platforms. The best solution for publishing in most non-English languages is to use the Unicode encoding, which is designed to support virtually all written languages.² This will solve the problem of storage, but for display, the user will still require a font that supports Unicode and that contains glyphs for the language to be represented. Unicode support is quite good in the latest operating systems, but those with older systems may encounter display problems.

inline objects

Documents that include other types of media (especially images) are very common. There are essentially two ways in which this issue may be handled: other digital objects may be embedded in the document (this is typically the way word processing software and PDF handles other media) or they may be linked and imported when the document is displayed (this is the method used in HTML, for example). Each method has its advantages and disadvantages. Documents with embedded files remain single documents, and are thus easier to manage, but they also tend to be much larger. Documents that link to other files are smaller, but the external files must always be available when they are displayed, so the management issues become more complex.

3) *What text-creation capabilities do I possess or have access to?*

The answers to this question depend on four factors: time, money, technology, and skill. The time factor depends on the project's deadlines. Since some methods of creating documents require more time to learn than others, a close deadline may obviate some of the available options. With funding to hire people with prerequisite skills, it may be possible to reduce the amount of time it takes to complete the project. Some formats and publishing environments require specialized technology, which in turn may require special skills and / or money. The following list describes some of the relative capabilities and costs associated with various text formats.

plain text

- easy to learn
- editors are cheap or free
- no formatting

word processors

- easy to learn
- software ranges from free to moderately expensive

² See <http://unicode.org> for details.

DRAFT

- large or complex documents may be more difficult to produce
- most formats are proprietary

HTML

- somewhat harder to learn
- HTML is just text, so any text editor can be used to produce it.
- HTML editors range from free to moderately expensive
- designed for display in a web browser, may not print out well

desktop publishing software

- requires some degree of skill or training
- moderately to very expensive
- will handle very large / very complex documents
- most formats are proprietary

XML

- requires some degree of skill or training
- Like HTML, XML is plain text, so any text editor can be used.
- XML editors range from free to moderately expensive
- can handle very large / very complex projects

SGML

- requires some degree of skill or training (more than XML)
- Like HTML, SGML is plain text, so any text editor can be used.
- editors range from free to very expensive
- can handle very large / very complex projects

4) *How static are the documents in the collection and the collection itself?*

Question #4 deals with the ongoing management of text collections. If documents in a collection will need to be updated periodically, then it will be helpful if there are organized ways of handling the updates. The management issue becomes more pressing the larger and more complex the collection of texts is. There may be indexes and tables of contents that need to be updated if new documents are added, or if documents change substantially. If documents contain links to one another, these may need to be updated also. If documents are to be updated, there may need to be a system in place for handling the different versions of the documents. If older versions should remain available, there will need to be a system in place for managing this. Even if older versions can be discarded, it may be desirable to keep track of what changes have been made and by whom.

5) *How sustainable should the collection be?*

Sustainability is another important issue. How long is the collection expected to be available in digital form? If the answer is "indefinitely," then it will be well worth considering the use of non-proprietary formats. If formats like MS Word or WordPerfect are used, it is likely that they will need to be migrated every few years to whatever format is in current use. This will be another ongoing management issue. Even the ongoing viability of HTML is by no means guaranteed. The formats most likely to be indefinitely sustainable are plain text, XML and SGML.

6) *To what kinds of uses might the collection be put in the future?*

The last question deals with the flexibility that may be required of your texts by future users. If your documents have been marked up in XML or SGML, and if the source versions are available, then other people will be able to query those documents in ways that you may not have thought of. This type of document allows for operations like semantic searching within documents or for linking

DRAFT

placenames to a Geographic Information System. Thus, while they are harder to produce, marked up documents will generally produce a much bigger payoff.

Formats

plain text

Plain text has many advantages as an archival and presentational format, particularly if it employs only the ASCII encoding, which has 128 character points, the first 32 of which are used for various control characters (carriage return, for example). ASCII Documents are likely to be displayable on any platform / software combination. In addition it is easy and cheap to produce. There are a number of excellent, free or inexpensive text editors available. The only real disadvantage of the format is that you sacrifice all formatting and structure (apart from line breaks and indenting) and give up the ability to embed other digital objects into your documents.

A (by no means exhaustive) list of editors:

- Emacs (free; unix/linux, windows)
- jEdit (free; any platform that supports Java)
- Notepad (free; Windows)
- vi (free; unix/linux)
- UltraEdit (shareware; Windows)
- BBEEdit / BBEEditLite (cheap / free; Macintosh)

word processors

Word processors are the most common tool used to create text documents. They are capable of producing documents with rich formatting and embedded digital objects (such as images). Microsoft Word dominates the category, and most other modern word processing packages will read Word files, though the reverse isn't necessarily true. Word is also capable of exporting to HTML, though the HTML it produces is rather idiosyncratic. Word processing software is typically very easy to learn. But word-processing formats do not typically support tagging the semantics of a document (e.g. distinguishing a personal or place name from other text). They may also have problems handling book size or very complex documents. In addition, their formats tend to be binary and proprietary, and are thus not easily parsed by other software.

Common word processors include:

- Microsoft Word (Windows, Macintosh)
- WordPerfect (Windows, Macintosh)
- Open Office / Star Office Writer (Windows, Linux, Macintosh – under development)
- Adobe FrameMaker (Windows, Macintosh, Unix, Linux – not v. 7)

HTML

HyperText Markup Language uses plain text with a system of elements or tags to describe the structure and formatting of documents. It includes mechanisms for the display of inline images or other media, but these are not embedded in an HTML file, rather they are linked to by means of a tag. HTML is actually an application of SGML. It is designed specifically for display in a web browser it does not necessarily make for good printed copy. Common HTML Editors include:

- Macromedia DreamWeaver
- Microsoft FrontPage
- Netscape Composer
- Macromedia HomeSite

PDF / PostScript

These are formats designed to be printable, and to preserve any and all formatting that the original document possessed. They are generally not formats that one authors documents in

DRAFT

directly, instead they are used as display or transport formats for documents authored in other systems. One advantage of PDF over PostScript is that it is searchable and can be handled by some screen-reading programs.

desktop publishing

Desktop publishing programs, such as LaTeX are designed for the production of camera-ready documents for print publication. They tend to be more robust (and sometimes more expensive) than ordinary word-processing software.

- TeX/LaTeX
- Adobe Pagemaker, FrameMaker, InDesign
- Quark Express

XML / SGML

Like HTML, SGML and XML documents are plain text, and any need for embedded objects is handled by links to external files. The main difference from HTML is that both SGML and XML prescribe the syntax for creating markup languages (like HTML), but do not prescribe particular tags to be used. Both provide a system, called a Document Type Definition (DTD) for declaring what tags may be used in a document, and how those tags are to be employed. XML is a subset of SGML, with more restrictive syntax rules, which allows it to be more easily parsed by computer programs. For this reason, there is much more software available for manipulating XML than SGML, and more of it is cheap or free. The great advantage of markup languages over other formats is that a well-designed markup language is capable of capturing semantic information in a text, as well as representing its structure and format.

There are several well-established DTDs available for marking up documents of various types. For works of literature, for example, the Text Encoding Initiative (<http://www.tei-c.org>) provides guidelines and a very mature and complete set of DTD documents with software for building a customized, TEI-conformant DTD for any project. The TEI is the standard format for SGML and XML document publication in the humanities. The TEI Consortium also provides consultation and training for projects wishing to use TEI. Other specialized schemas include MathML, Chemical Markup Language, Physics Markup Language, and Geography Markup Language. Projects that are creating large and/or complex documents should seriously consider using or extending an established standard, rather than creating one from scratch.³ Editors include:

- Emacs (free; SGML and XML; Unix, Windows)
- XMLSpy (XML, Windows)
- XMetal (XML, Windows)
- FrameMaker 7 (SGML and XML, Windows, Macintosh, Unix)
- jEdit (free; XML, any Java-enabled platform)
- TurboXML (XML, any Java-enabled platform)

Images

It is, of course, possible to deliver texts as images. This may seem a particularly attractive option if one wishes to scan a printed text and display it without doing any extra work. The real problem with this approach is that it produces text that is totally inaccessible to visually impaired people who rely on screen readers. We therefore do not recommend that texts be handled in this fashion. If it is necessary to scan printed texts, the text should be run through an Optical Character Recognition (OCR) program and corrected.

³ [A list of XML applications may be found at http://xml.coverpages.org/xmlApplications.html.](http://xml.coverpages.org/xmlApplications.html)

Web Display

Modern web browsers can handle most text formats. ASCII (plain text) text files require the least amount of space (one byte per character), but they cannot handle any advanced formatting, including italicization and bold marking. Rich Text Format (RTF) allows for this formatting, but not much more. Microsoft Word documents and HTML documents allow for advanced features including internal and external hypertext referencing. And Adobe PDF is useful for publication but not for shared editing of documents. The most innovative solution is the use of XML documents, which can be rendered into any of the other formats mentioned below using XSL stylesheets, or they can be displayed directly in some browsers with an XSL or CSS stylesheet. Knowing who you are publishing for and why will help you determine the most appropriate format.

Text for web display

Format	Adobe PDF; Word (or other word processor); RTF; XML+stylesheet; HTML; ASCII/Unicode(TXT)
Campus Help	CIT, CTL, ITRC, OASIS, UNC Libraries
Recommended Editors	Notepad, Wordpad, Word, Adobe Acrobat; XMLSpy, JEdit, Xmetal, OpenOffice Writer, Adobe FrameMaker, Macromedia HomeSite, Macromedia Dreamweaver
Recommended Viewers	Web browser, Acrobat Reader, see above

Printing

For most text documents, you will probably find that the formats mentioned under Web Display (above) are adequate.

There are, however, other formats that allow for more control over layout and printer-specific commands. PostScript (PS) and Encapsulated PostScript (EPS) are both such formats, as well as LaTeX, which even has its own language used for formatting and rendering text documents. XML can be rendered into a number of print-ready formats. Unfortunately, all of these print-specific methods require some expertise to execute correctly.

If you are readying text for publication in a journal, book, magazine, etc., you should contact your publisher and inquire about its preferred specifications.

Text for printing

Format	(Encapsulated) PostScript; XML+XSLFO; LaTeX
Size	150-300 dpi
Campus Help	CIT, CTL, ITRC, OASIS, UNC Libraries
Recommended Editors	Adobe FrameMaker; TeX/LaTeX; OpenOffice Writer, Microsoft Word (for smaller documents), Microsoft Publisher, Adobe InDesign, Adobe Pagemaker, Adobe Framemaker

Archiving

If you think you will need to access a text document again over a longer period of time, it is recommended you keep an archive version. How you archive text depends directly on what the text is intended to do. Certainly if you're writing a simple document without any necessary formatting,

DRAFT

archiving text as a simple TXT document, edited in something as simple as Notepad, is fine. But as documents get more complex, both in contents and formatting, it is important to archive as much information as possible. We recommend using XML along with a document type definition (DTD) or schema, which will allow you to define your document's structure and can be used to validate your final document. Because XML is both standardized and human-readable, encoding your document in this manner will ensure its longevity even as the editing programs and formats themselves change. For archiving text, we strongly recommend avoiding proprietary formats such as Microsoft Word.

Text for archiving

Format	RTF; ASCII/Unicode(TXT); XML+DTD (e.g., MathML, TEI, DocBook, etc.); SGML+DTD; LaTeX; (E)PS OASIS
Campus Help Recommended Editors	Notepad, Wordpad, XMLSpy, JEdit, XMetal, OpenOffice Writer, any other text editor or SGML/XML editor that does not save to a proprietary format.
Recommended Viewers	Any XML/XSL compatible Web browser

Presentations

Because of the CCI Microsoft agreement, for presentations to be given on campus, the obvious choice of software is Microsoft PowerPoint. If you want to share presentations with people who might not have PowerPoint, or you want to present over the web, there are a number of alternatives: Internet presentation emergent standards such as SVG (Scalable Vector Graphics) and SMIL (Synchronized Multimedia Integration Language) offers universal compatibility, including accessibility for disabled users.⁴ Also, for those more comfortable with Internet multimedia programs, Macromedia's Flash not only creates interactive presentations but saves into standard formats. It is most important to retain the original file, whatever format you prefer to create; while the standards of the output formats will not change, *how* they are saved from a particular editor is changing rapidly as the manufacturers continue to align their programs with these standards.

Presentations

Format	Microsoft PowerPoint; Macromedia Flash; Adobe Persuasion; SVG; SMIL
Campus Help	CIT, ATN Training, ATN Help Desk
Recommended Editors	Microsoft PowerPoint; Macromedia Flash; Adobe LiveMotion; GRiNS (SMIL)
Requirements	Web browser (and plug-in)

⁴ While these standards are being adopted and promoted widely, they still require special programs and/or browser plug-ins to work properly. These programs can be found at the W3 Consortium's site (<http://www.w3.org/>).

Other

Computer Code. Similar to the reasons for retaining both original and output versions of presentations, it is important to retain the original computer code alongside its binary derivative.

Data Sets. Our only recommendation for data sets is again similar to that above: save the original format of the data if at all possible. Because the data sets used on campus vary widely both in origin and purpose, no one program can be recommended for handling them. There are, however, “expert pockets” on campus which should be consulted when working with particular data sets: e.g., the Odum Institute for Research in Social Science, the Department of Geography or the Ancient World Mapping Center for geographical information systems (GIS) data, the School of Information and Library Science for bioinformatics data, etc. This document defers to these groups as experts in these areas and will not discuss them any further here.

Storage

Storage options for your digitized media should be considered *before* you begin digitizing. Storage space needs vary significantly, depending on file formats and the quality of media desired. Backup policies should always be implemented.

There are numerous solutions for storing media. Computer hard drives, ZIP drives, and CD or DVD offer local but limited storage space for many media types. However, all these media degrade to various degrees over time. If you wish to use a CD or DVD burner, you must analyze your data and decide what kind of storage media suits you best. CDRs hold only 700 MB, DVDRs can hold up to 4.7 GB. For video projects, DVDR is the only real option; for any other media type, CDR will suffice.

For storage of larger media, particularly raw digital audio and video, you might consider an external hard drive. There are many on the market up to 120GB which utilize the new Firewire standard for faster access, meaning that you can do your video capturing and editing directly on this drive. DAT is also a commonly used medium for storing audio, and miniDV/Dvcam for video.

If your media are directly related to a particular department or project, you can store the media in networked space (AFS) that can be accessed from remote locations. Projects are currently allotted up to 10GB per project or department, with more available to purchase. Of course, you may also use your own personal AFS space, currently limited to 250MB.

If you have a fairly large number of media files or expect to build a collection over time, UNC Digital Library Services may be an appropriate option for storing, managing and sharing your collection. Primary support at this time is for collections used to support instruction, but this new system is designed for research collections as well. For more information on the project and who to contact with questions, see the project home page at www.unc.edu/projects/diglib/, or send email to diglib@unc.edu.

ATN also offers university projects a remote tape backup system called "mass storage," which allows for a more permanent method archiving. It is not, however, as quickly and easily accessible as local media and thus should be used more as a strict archive than a working space. Additional information on this service is available at: http://www.unc.edu/atn/mass_storage/

Looking for Something Else?

If you have multimedia needs that are not addressed in this document, send email to mediasupport@unc.edu, or contact a campus service organization directly. See Campus Help below.

Campus Help

General Digitization Support

Academic Affairs Library

- Media Resources Center (<http://www.lib.unc.edu/house/mrc/index.php>)
- Collaboratories (<http://www.lib.unc.edu/house/index.php?display=collaboratories>)
- Photographic Services Section (<http://www.lib.unc.edu/ncc/copies.html>)

Phone: 962-1355

Email: kimv@email.unc.edu

Center for Teaching and Learning (instructors only)

Teaching Resource Lab (<http://ctl.unc.edu/csssml.html>)

Phone: 966-1289

Email: ctl_unc@unc.edu

Center for Instructional Technology (instructors only)

- Audio/Video Services (<http://www.unc.edu/cit/vidserv/index.html>)
- Other services (<http://www.unc.edu/cit>)

Phone: 962-6042

Email: cit@unc.edu

College of Arts and Sciences (College-affiliated faculty/instructors only)

OASIS Media Services (http://oasis.unc.edu/services/media_services.html)

Phone: 843-2205

Email: diglib_media@unc.edu

Health Sciences Library

Media Kitchen (<http://www.hsl.unc.edu/mk/MKslices.htm>)

Phone: 962-0800

IT Response Center

<http://help.unc.edu>

Phone: 962-HELP

Other Campus Resources

The following campus services, projects and repositories may also serve as valuable references for your digitization goals:

- Ancient World Mapping Center
<http://www.unc.edu/depts/awmc/>
- Documenting the American South
<http://docsouth.unc.edu/index.html>
- Ibiblio
<http://ibiblio.org/>

DRAFT

- Medical Illustrations and Photography
<http://www.med.unc.edu/wrkunits/4serv/medill/>