3 Digitization

This chapter discusses processes of digitization, the methods for creating digital files from various source materials and turning them into standard formats. It raises points about the advantages and disadvantages of different formats and introduces basic techniques in making browser-readable files using HTML (Hyper Text Markup Language).

3a Digital documents: formats and protocols

Bonus content: copying code and using templates

Software programs for creating web-based materials have been around for decades. These programs make it possible to prototype web functionality, make design mockups and functional sites, and to put materials online that have features through a drag-and-drop or graphical design interface. They make it easier to create sites with advanced functionality, even without sophisticated skills, as do standard platforms like blog, wiki, WordPress, and others. The trade-off is that customizing and maintaining such projects can be difficult. The Cascading Style Sheets (CSS) and other components that support the functions may or may not be accessible. Software programs for web design often include code that will be hard to read and may be part of proprietary materials that are referenced by a link, but not accessible to the individual user.

Copying code from others is a questionable enterprise since it means taking the work someone else has done and appropriating it for one's own use. While new content may be put into the design, using someone else's code is still an act of appropriation—possibly even plagiarism. A good exercise is to see what happens in a classroom when students "borrow" code from each other—with and without permission or acknowledgment.

Exercise 3.1: View developer

Go to one of your favorite websites. Right-click on the webpage and view these same pages in the "View>Developer" or "View>Source" option in your browser. How many HTML tags do you recognize in the page when viewing it as code? You may not see any HTML tags, just a pointer to a library of materials on a server. How do you understand that as scaffolding for a webpage?

Exercise 3.2: W3schools HTML tutorial and standards

Work through this HTML tutorial (http://www.w3schools.com/html/) and its exercises. Learn more about styling your HTML with cascading style sheets by working through these CSS exercises (https://www.w3schools.com/css/default.asp). Look at W3C standards (https://www.w3.org/TR/2021/NOTE-html53-20210128/) to see how recently they were updated.

Based on what you learned through the W3schools tutorials, use a text editor, like Visual Studio Code (https://code.visualstudio.com/), and build a small personal webpage that introduces you and a few of your favorite things. It should include the following elements and be sure you understand what each of the elements is:

- A document type declaration
- Head (<head>) and body (<body>) sections
- A style section with CSS for three different elements that adjusts colors, fonts, and borders
- Two different header texts (such as h2 and h3)
- A paragraph text
- An image
- A link
- For special characters, please consult these references for HTML symbols (https://www.w3schools.com/html/html_symbols.asp) and HTML UTF-8 diacritical marks (https://www.w3schools.com/CHARSETS/ref_utf_diacritical.asp). This is particularly important for math and foreign languages.

Exercise 3.3: Deprecated tags JavaScript examples

Look at these deprecated HTML tags (http://www.w3docs.com/learn-html/deprecated-html-tags.html) and see if you can figure out why they were abandoned.

How-to example

The HTML tags listed on that page were abandoned primarily due to advancements in web design standards, especially with the introduction of CSS and HTML5. Older tags like <font>, <center>, and <marquee> were replaced by CSS for greater flexibility and separation of content from design. Some tags were removed because they either provided redundant functionality or were too browser-dependent, making them unreliable. More modern alternatives were developed to ensure cleaner, more maintainable, and accessible code.

Consult the most up-to-date tag set standards (https://www.w3schools.com/TAGs/).

Exercise 3.4: JavaScript examples

Look at these JavaScript examples (http://www.w3schools.com/js/default.asp) to see if you can read them.

Return to the personal web page you created in Exercise 3.2. Pick JavaScript features to add to that HTML document to bring an element of interactivity.

Exercise 3.5: Guidelines for accessibility

Look at these Web Content Accessibility Guidelines (http://www.w3.org/TR/WCAG20/) and consider steps for implementation. Also, try out this ADA testing 101 (https://webflow.com/blog/ada-testing) feature to see if it helps your design.

Recommended readings

"An Introduction to Multilingual Web Addresses." *World Wide Web Consortium (W3C)*, Accessed April 19, 2025. www.w3.org/International/articles/idn-and-iri/.

Doty, Nick, Alice Cooper, and Wendt Seltzer. 2023. "Human Rights and Technical Standard-Setting for the Web." *Center for Democracy and Technology*. https://www.ohchr.org/sites/default/files/documents/issues/digitalage/cfis/tech-standards/subm-standard-setting-digital-space-new-technologies-standard-setting-organizations-w3c-participants-7-input.pdf

Raggett, David, Jenny Lam, Ian Alexander, and Michael Kmiec. 1998. "A History of HTML." *Raggett on HTML 4*, Addison Wesley Longman. www.w3.org/People/Raggett/book4/ch02.html.

Webflow team. 2023. "Web Standards: Where They Came from and Why They Exist." https://webflow.com/blog/web-standards

Zhang, Sarah. 2016. "Chinese Characters Are Futuristic and the Alphabet is Old News." *The Atlantic*. www.theatlantic.com/technology/archive/2016/11/chinese-computers/504851/.

3b Digitization and file formats

Bonus content: file naming, folders, and organization of project materials

File naming is an art and requires careful planning if it is to be useful. Filenames should be readable by humans as well as computers whenever possible. They will be used to sort as well as retrieve information. Once file names and folder structures are established for a project, they will be very difficult to alter without introducing errors into tags and pathways. To make files human-readable, a naming convention should be established for the entire project at the outset that relates to the conceptual framework of the research. The hierarchy of organization should match the conceptual model of the project. If you are a historian studying maps of battles in a particular region, should the top-level files be "Battles," with "Maps" as an element? Or should the "Maps" be the highest level because a map is used to show multiple battles? Should books be organized by authors or publishers? Dances by choreographers or individual performances or dancers? Films by directors or studios? Artifacts to cultures or collections? These questions quickly shift from abstractions to contested territory as part of the conceptual design of a project.

These decisions should be thought through at the beginning, since reworking file structures is expensive, time-consuming and leads to errors. In general, to be computer-readable, file names should not contain any "special" characters (e.g. #,!, or.and other "reserved" signs). They also should not contain spaces. So, your file named "My House, Pictures of exterior" should be "My_House_Pictures_Exterior" or "MyHousePicturesExterior."

Decisions about the basic organization and structure of files and materials can be governed by professional bibliographical conventions. Archivists, museum curators, and other information practitioners have standards for this kind of organizational work, and where these exist for a field, they should be followed. When these do not exist, then conceptualizing a rational framework of organization is crucial and is also part of the intellectual argument of the research.

Bonus content: checklist for digitization of different media formats

Analog texts

- Digitize texts by scanning, photography, or re-keyboarding.
- OCR-readable scans allow the text to be searched (some errors may occur).
- Photographic facsimiles of pages retain format features but will lack the capacity for search.
- Typing (keyboarding) a text produces a clean, ASCII-based file that can be analyzed and data mined, but is time-consuming and can introduce errors.

Images

- Determine the resolution to be used for scanning as well as the file format. 72 dpi works for screen resolution and is efficient for web storage and display, but not adequate for print.
- Consider the purpose of the project (image analysis, research, web presentation) and determine which is the appropriate format: JPEG, TIF, RAW, or other standard.
- Create high-resolution scans at 600 dpi in TIFF format to create files from which other derivatives will be produced (thumbnails, screen display, or print quality of about 300 dpi).
- Understand the difference between color formats. Many print images are created in CMYK color space (pigment-based) while screens display in RGB (light-based). A translation of these color formats is critical for certain kinds of work.
- Examine the original format and medium of the image and determine which features are crucial for the research work to determine file format decision.
- Photograph original works of art or scan them with large-format devices. Color correction, lighting, and other elements are crucial.

Sound recordings

- Analyze the original recording and its features to determine what should be preserved in the digitization (bandwidth, dynamic range, frequency, noise, etc.).
- Use a microphone to re-record audio by converting air vibrations to digital information.
- Use a direct transfer method.
- Consider whether a WAV file, which is very high resolution, or a lower resolution file like MP3 is more suited to your purposes.

Video or film

- Use video and film converters for direct transfer of time-based media into digital files; playback and frame rate are important considerations when transferring media.
- Consider projecting and re-recording film using a digital camera.
- Capture analog video through a USB attachment from a tape deck and the correct software.
- Transfer digital video to copy it directly.

3D objects

- Scan, video-record, or photograph the object.
- Provide a 3D representation of an object that can be viewed from all sides by using photogrammetry.

AR, VR, and XR projects

- Select a platform for your project (Unity is a good starting point, especially for beginners).
- Start your new project in the program and configure the various plug-ins.
- Determine what functionality you want and which assets you need to build this into the project.

Exercise 3.6: Format test

Using a scanner, your phone, or other device, photograph or scan an analog image. Put it into a TIF file and a JPEG file. What are the differences in file size? Look at the file in an enlarged form—try zooming in and out. What differences are there in appearance? Why does it matter?

How-to example

Responses to the questions in Exercise 3.6 should demonstrate a basic technical understanding of file formats and compression, critical observation skills when comparing the visual and size differences, and contextual thinking about when and why to use each file type based on the needs of a project or task.

Answers would likely address:

- File size differences
  - TIFF file: Likely much larger due to being an uncompressed or lossless format.
  - JPEG file: Smaller because it uses lossy compression to reduce file size.
- Image quality differences (zooming in and out)
  - TIFF file: Likely to maintain better detail when zoomed in due to its higher quality and lack of compression artifacts.
  - JPEG file: May show pixelation, compression artifacts (blocky areas or fuzziness) when zoomed in, especially in areas with subtle color gradients or sharp edges.
- Why it matters
  - TIFF: Ideal for high-quality archival purposes, printing, or professional image editing where detail and accuracy are critical.
  - JPEG: Suitable for web use, sharing, or when storage space is limited, but less suited for tasks where image fidelity is crucial.

Exercise 3.7: Quality control

Do a web search for images of the Mona Lisa and specify size. Compare the images. Why are they so different in color? In detail? Even in cropping? What does this tell you about making digital files from analog materials?

How-to example

Results from this type of image search will show how technical aspects of digitizing, such as resolution, file size, compression, and post-processing, affect image quality.

Responses to the questions posed in Exercise 3.7 should demonstrate. The variability in digital files has broader implications for how we digitize and preserve analog materials. For example, note the file sizes, which can range from small thumbnails to high-resolution images. Notice that larger file sizes often correlate with better detail or resolution, while smaller images may be pixelated or blurry when enlarged. Identify instances of color variation across images, with some appearing more vibrant, while others are more muted or yellowed, and consider why these color differences might be. It could be due to differences in lighting when the photo was taken, camera settings, or post-processing techniques used to adjust color. Observe that some images have greater clarity and sharpness in the details, while others might appear blurry or over-smoothed. Difference in detail could be attributed to the quality of the scan or photo, the resolution, and any compression applied to the digital file. Also, note that some images of the Mona Lisa are cropped differently, with some focusing only on her face, while others include more of the background or even parts of the frame. This variation could be due to choices made by the person digitizing the image, deciding which part of the artwork to emphasize or how to best fit the image into certain formats or platforms. Differences result from the choices made during the digitization process—camera settings, lighting, post-production editing, file format, and compression techniques all affect the final image. Recognize that digitization introduces variables that can affect the accuracy and fidelity of the final digital representation of the analog material.

Recommended readings

Analog to digital conversion, Geeks-for-Geeks. 2024. https://www.geeksforgeeks.org/analog-to-digital-conversion/.

Manzuch, Zinaida. 2017. "Ethical Issues in Digitization of Cultural Heritage." *Journal of Contemporary Archival Studies* 4. https://elischolar.library.yale.edu/jcas/vol4/iss2/4/.

Rouhani, Bijan. 2023. Ethically Digital: Contested Cultural Heritage in Digital Context. *Studies in Digital Heritage.* https://scholarworks.iu.edu/journals/index.php/sdh/article/view/35741.

Bibliography

Choi, Charles Q. 2020. "World's First Classical Chinese Programming Language— IEEE Spectrum." *IEEE Spectrum: Technology, Engineering, and Science News*. spectrum.ieee.org, https://spectrum.ieee.org/tech-talk/computing/software/classical-chinese.

Goldsmith, Jack. 2018. "The Failure of Internet Freedom." In *Knight First Amendment Institute*. New York: Columbia University. https://knightcolumbia.org/content/failure-internet-freedom.

Noble, Safiya. 2018. *Algorithms of Oppression*. New York: New York University Press.

Rumsey, Abby Smith. 2016. *When We Are No More: How Digital Memory Is Shaping Our Future*. New York: Bloomsbury Press.

Simpkin, Sarah. 2020. "Getting Started with Markdown." https://programminghistorian.org/en/lessons/getting-started-with-markdown.

Resources

- Data best practices and File Formats, Stanford University (https://guides.library.stanford.edu/data-best-practices/format-files)
- File naming best practices, Harvard University Library (https://guides.library.harvard.edu/c.php?g=1033502&p=7496710)
- JPEG standard (https://jpeg.org/about.html)
- Open Source Formats (https://en.wikipedia.org/wiki/List_of_open_formats)
- Packet delivery (http://www.sciencedirect.com/topics/engineering/packet-networks)
- Sound recording digitization (http://digitalsoundandmusic.com/5-1-2-digitization/)
- W3C Consortium Mission Statement (http://www.w3.org/Consortium/mission)
- WAVE guidelines (https://wave.webaim.org/)